Multimodal Language Processing Using NLP Approaches

Fernando Ferri, Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche, via Nizza 128, 00198 Rome, Italy; E-mail: fernando.ferri@irpps.cnr.it

Patrizia Grifoni, Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche, via Nizza 128, 00198 Rome, Italy; E-mail: patrizia.grifoni@irpps.cnr.it

Manuela Tersigni, Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche, via Nizza 128, 00198 Rome, Italy; E-mail: Manuela.tersigni@irpps.cnr.it

ABSTRACT

People usually communicate through multimodal dialogue. Multimodal interaction is in fact flexible and natural because it uses all five senses in parallel. For this reason we need to consider multimodal language definition and processing adopting the techniques and approaches used in Natural Language Processing (NLP). We describe the characteristics of a multimodal language by NLP, considering that the speech mode appears to be the most complete (it is considered the predominant mode). Users communicate and interact through reference to a set of key concepts. These can be expressed with different modes and/or by more than one mode simultaneously. When defining a multimodal language, these key concepts must be extracted. They are then processed using a natural language approach: any concept expressed in any mode can be "translated" into natural language. This implies that speech acts as a "ground layer" that all the modes refer to. We propose a tool to define multimodal languages, which allows the user to define the language in his/her own way to express concepts of a particular domain in the different modes.

Keywords: Multimodality, multimodal language, fusion, ambiguity.

1. INTRODUCTION

The purpose of this paper is to identify characteristics of multimodal languages in order to design a tool that enables the user to define them in some specific contexts. For instance, a possible case is the use of formal graphic models such as Unified Modelling Language (UML) and Entity-Relationship (E-R) diagrams during a project meeting. A second case is the use of tourist maps with which the user can interact to ask for information on restaurants, museums, theatre schedules, transport and so on through multimodal interaction on mobile devices.

A multimodal language is a language that allows people to communicate with a system synergistically through multiple modes (i.e. speech, sketch and writing).

In several contexts, such as those cited above, speech is the most complete (predominant) mode (users tend to explain everything orally, using other modes to support what they say). By predominant mode we mean the mode the system first refers to. It considers other modes only if it needs to solve any ambiguous or incomplete cases.

For this reason we need to consider the definition and processing of multimodal languages according to the techniques and approaches used in Natural Language Processing (NLP).

The study of a synergistic system cannot be separated from the study of the way a machine processes the natural language, i.e. Natural Language Processing (NLP). This presents many problems, the biggest of which is language ambiguity. Oviatt et al [1] explored whether a multimodal architecture can support *mutual disambiguation* (MD) of input signals. Mutual disambiguation enables recovery after unimodal recognition errors, leading to a more stable and robust performance, as it permits the strengths of each mode to overcome weaknesses in the others [2].

Oviatt [3] also highlighted the differences between unimodal and multimodal communication with respect to the structure of spoken language. Sharon Oviatt & Karen Kuhn showed that multimodal language differs from spoken language in its brevity, semantic content, syntactic complexity, word order, disfluency rate, degree of ambiguity, referring expressions, specification of determiners, anaphora, deixis and linguistic indirectness.

Qiaohui Zhang et al [4] handle ambiguity by using gaze information to integrate the user's speech input. If multiple objects are chosen simultaneously due to an ambiguous description, the one closest to the gaze fixation will be the multimodal result.

The ambiguity issue is addressed in [5] by designing a multimodal agent for route construction (MARCO).

This paper discusses both NL disambiguation and multimodal user interfaces. [6] presents studies on the combined use of different input modes. A component-based approach to specify and develop multimodal interfaces using a mode-independent fusion mechanism is described in [7].Michael Johnston [8] describes a multimodal language processing architecture in a unification-based grammar formalism.

Sections 2 and 3 summarise our approach to and problems related with fusion at the multimodal language level and the natural language processing approach in a multimodal environment. Sections 4 and 5 describe the system to define the multimodal language and in section 6 we draw some conclusions.

2. FUSION AT THE LANGUAGE LEVEL AND DEFINITION OF THE MULTIMODAL LANGUAGE

The synergic integration of the system's various input channels can be achieved in several ways. One possibility is to consider the input channels separately and then merge them [12]. An alternative approach is to carry out the integration during the definition of the multimodal language. This requires the capture of its characteristics using natural language and an understanding of the intrinsic nature of natural language; i.e. rebuilding of the language structure and consideration of how the same concept can be expressed by different input types and how implicit references (deictic expressions) can be solved. A deictic expression refers to the personal, temporal, or spatial aspect of an utterance; its meaning therefore depends on the context in which it is used. Examples of deictic expressions are "this", "that", "here" and "there".

When defining a multimodal language some key concepts for a particular domain are extracted. These can be then expressed in different modes, but are processed with a natural language approach. Any concept expressed in any mode can be "translated" into natural language. This implies that natural language acts as a "ground layer" that all the modes refer to, making speech the predominant mode and the key to the fusion of different modes.

Our study therefore involves NLP to understand how to obtain information from a text analysis and exploit it alongside other information conveyed by other modes, thus demonstrating the approach to solve such problems in a multimodal environment.

Through this approach, the definition of the multimodal language can be divided in two phases: the first is parsing-driven, the second oriented to tangible representations of linguistic experiences.

In the first phase some key concepts (expressed according to the various modes) are identified. They are encapsulated in structures (templates), which constitute their "frame" and define the language giving semantic value to what the user says or draws. A template is a syntactic structure consisting of concepts (expressed in the various modes) and syntactic categories, which are assigned with a given semantic value.

Each user action can match one or more templates. However, the multimodal language is not produced from a rules set but from a sentence analysis set - it is deduced from the spoken language. This helps to locate and eliminate syntactic ambiguity, as admissible syntactic structures (at the parsing level) that do not belong to the admitted structures set are not considered.

3. NATURAL LANGUAGE PROCESSING AND MULTIMODALITY

The first step we took was to discover how to pick up information produced by the user from the speech channel.

The study of NLP is important because we use natural language as the basis of our approach to fusion: the speech mode is predominant, while the others are used as a support when needed.

Most human communication is through speech, though in some cases the use of other modes makes it simpler to understand one another and convey concepts: this improves synthesis and precision, especially when relevant concepts need to be communicated.

NLP is thus the starting point for fusion. However, understanding a language involves - among other things - knowing what concepts a word or a sentence stands for and how they are related. The use of other modes therefore plays an important role in helping to solve problems that often arise in natural language comprehension.

In the following section the levels of NL understanding are discussed and some problems related to NLP are described from a multimodal point of view, i.e. with the help of other input channels.

3.1 Levels of Natural Language Understanding

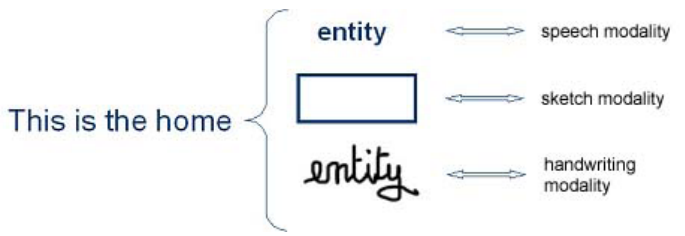
To process natural language a machine has to receive inputs from other modes and have a comprehension of natural language at different levels.

To determine what a user is saying, the system has to analyse an incoming audio signal and recover the exact sequence of words that the user used to produce that signal. This task requires knowledge of phonetics and phonology. The system also has to recognise word variations and contractions (morphology). Furthermore, a syntactic knowledge of the language is necessary to understand which word order makes sense. To this end, we used a statistical lexicalised parser that effectively solves the ambiguity problem. How this is done is explained below.

Understanding the nature of a request requires knowledge of the meaning of the words making up the sentence (lexical semantics) and the ways they can be combined to make a meaningful sentence (compositional semantics).

At this level, we used the information conveyed by the other modes: the key point is that the same concept can be expressed in a number of ways, which must be considered as semantically equivalent. This assigns a semantic value to the user's actions. The semantic side is therefore handled at different levels: first, at the level of each single mode, then at the multimodal level, when the modes are considered synergistically.

Figure 1. Example of the definition of an association in the E-R diagram scenario



The resulting multimodal language consists of these associations among different ways of expressing the same concept. This allows the problem dealt with at the fusion level to be solved at a language level. For example, in Figure 1, the sentence:

"this is the home entity"

has the same meaning as saying:

"this is the home" while drawing a rectangle,

once the entity concept and the rectangle shape have been associated.

A higher level of natural language understanding is founded on pragmatics, i.e. the knowledge of how words are used in everyday life to make conversation easier. This level is dealt with through deictic expression handling.

Some problems which arise in NLP are described below. We will see how the synergistic use of other modes supporting speech can help solve them.

Ambiguity is one of the biggest problems for NLP. First, it increases the range of possible interpretations of NL, and a computer has to find a way to deal with this. There are various types of ambiguity. These include category ambiguity, in which there are a number of grammatical terminal symbols for the same word; for instance, the word "time" can be both a noun and a verb. This can sometimes be resolved by syntactic analysis.

Another type of ambiguity is related to the meaning of a word, which may correspond to only one terminal but a number of different concepts. In fact, many words have more than one meaning; we have to select the meaning that makes the most sense in the context. Temporal observations are essential for this purpose: if a word (speech mode) appears to be ambiguous, it is possible to examine the other modes to rebuild the user's original intention.

A third type of ambiguity is structural, and consists of the existence of more than one parsing of the same sentence. For instance, the sentence "choose between A and B or C" has two possible interpretations:

1. [A] and [B or C]
2. [A and B] or [C]

Referential ambiguity arises when a language does not specify to which word an adjective refers. For instance, the sentence: "pretty little girls' school" can have various interpretations: the school is small, the girls are small, the girls are pretty, the school is pretty.

The use of more than one mode can help to resolve the syntactic and semantic ambiguity in the speech mode: considering speech as predominant, the other modes are called in to clarify the meaning of any ambiguous or incomplete sentences. Any ambiguity in the speech can be disambiguated by examining the information provided by the other modes to obtain the sense of the sentence.

Some types of ambiguity arise at the very moment that other modes are introduced. These are described below:

Ambiguity Caused by Deictic Expressions

Even if deictic expressions (and references in general) can help directly identify what the user is referring to, they can also be a source of ambiguity. When a user pronounces a deictic expression, it is not always clear if s/he is referring to something said previously or something that s/he has drawn or is drawing. The problem gets even more complex in the common situation that the user draws a figure while explaining what s/he is drawing using deictic expressions that do not refer to the figure directly. For instance, if the user says: "And this is the solution to this problem" while drawing a symbol matching a concept, the system has to understand that only the first deictic expression refers to the drawn object, while the second refers only to the discourse, without involving any modes other than speech.

The risk that deictic expressions will create ambiguity in a multimodal system is also related to another factor - their vagueness. In fact, people are not used to specify precisely what a deictic expression refers to, and so the system often has

insufficient information to understand what the user means. The use of temporal windows can be of help: if the user draws an object, the deictic expressions pronounced in a certain temporal interval refer to such object.

Structural Ambiguity

This arises when different multimodal inputs overlap, causing an incorrect interpretation of the user's action.

Let us consider the following example: a user pronounces the sentence "This is the home". Suppose that at this point, the system expects either the concept of entity or that of relation (expressed in any mode). The user may draw the object corresponding to the concept of entity (a rectangle) and at the same time carry on expressing the concept of relation in another mode:

This is the home, *the relation now has to be found.*

If the sketch of the rectangle temporally overlaps the pronunciation of the word "relation", it is not clear if "house" is a relation or an entity.

4. THE SYSTEM ARCHITECTURE

Starting from NLP approaches according to a set of concepts in an application domain, we have designed and implemented a software tool, which provides users with the functions to define a multimodal language.

The system consists of two environments: in the first, the user can define the language in his/her own way to express concepts in the different modes (in our case speech and sketch); in the second, s/he can use the language to interact with the multimodal interface.

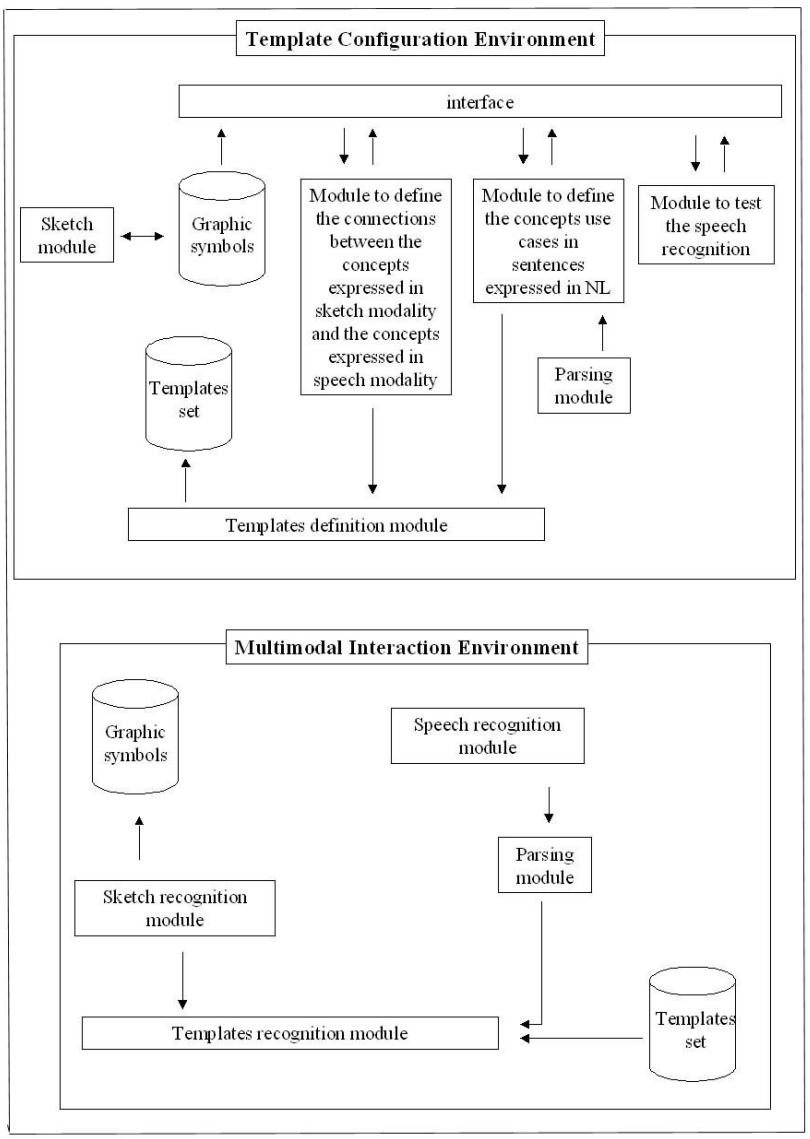
The definition environment is immersed in the fusion system, so that the user can switch from one environment to the other at any time.

The configuration environment has the aim of defining a set of structures - the templates - that enable the multimodal interaction to be processed.

Templates are defined through the following steps:

1. concept definition
   The user locates some concepts s/he thinks are relevant for his/her aim. For example in a E-R diagram definition, the user may locate the concept of entity.
2. correlation of concepts with "signs"
   Once the concepts have been defined, the user provides a sign to represent

Figure 2. The system architecture



these concepts in the speech and sketch modes. For example, the user may connect the concept of entity with the string "entity" (speech mode) and the rectangle (sketch mode).

3. correlation between signs
   The user is required to define a set of use cases (strategy "by example") for the relevant concepts in natural language. For example, the concept of entity is used to create a new entity, the user provides the example "this is the home entity".
4. template learning
   The system generalises the use cases provided in the previous step and builds the template set. It therefore:
   - assigns any deictic expression to the "deictic" category ,
   - replaces the concept with the related signs,
   - replaces any other word with its syntactic category.

In the example, the template obtained will be:

Deictic + VBZ + DT + NN + ("entity" or rectangle)

The template can thus be defined as the syntactic structure of an NL expression of a use case of the relevant concepts. It can then be used to interact with the multimodal system, in which the user can draw and speak at the same time.

A vocal recognition tool writes what the user says on a text editor, while a sketch recognition tool works on the user's drawing. The system makes the fusion between the two signals, recognising any matches between what the user said/drew and the template set. The two input streams are compared on temporally and according to the templates to see if they are complementary or redundant. For example, the multimodal dialogue may contain the following speech:

.....that, well, is now a new entity, the employer (rectangle)

The system architecture is summarised in figure 2: the configuration environment consists of an interface, which communicates with a graphical symbols set and with:

- the module to connect concepts with "signs",
- the module to provide use cases of the concept in NL,
- the module to test the template recognition algorithm.

These modules communicate with:

- the sketch module, which communicates with a graphic symbols library,
- the parsing module [13] which communicates with the module to define the use cases which define the template set,
- the module to define the templates, which communicates with the module to define the connections and the one to define the use cases. This module creates a set of generalisations of the use cases, i.e. a set of templates.

To define the template set, the use cases are provided as input to a natural language parser (in fact use cases are expressed in natural language).

Figure 3. The system

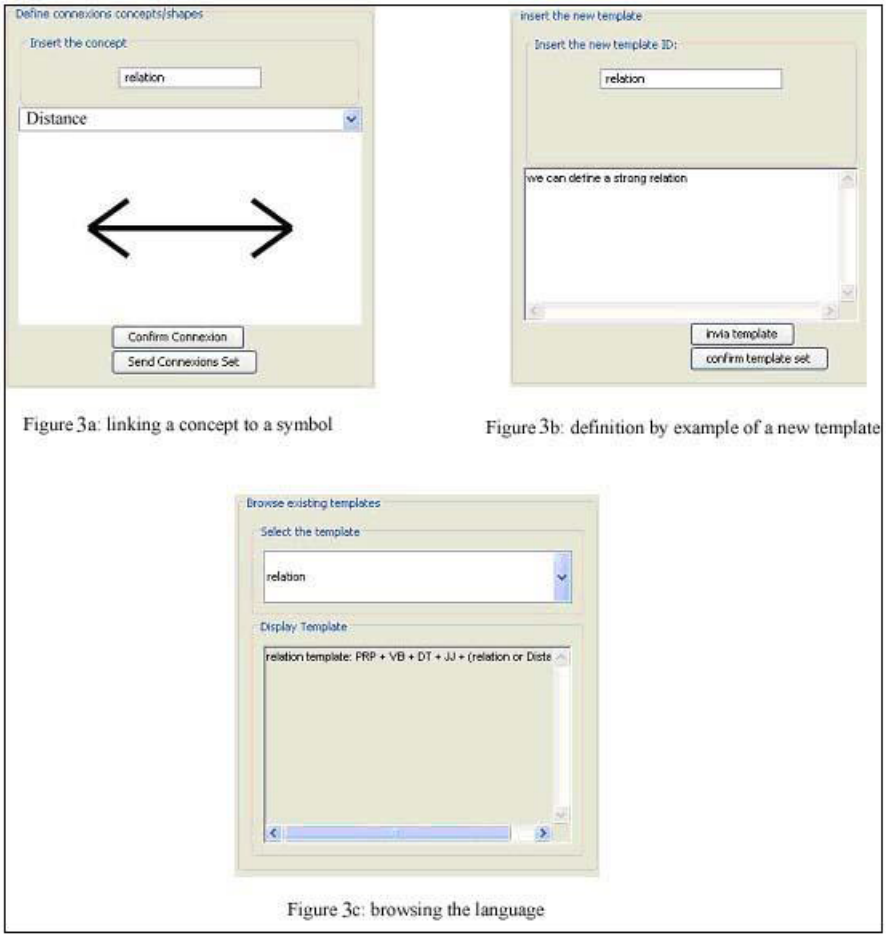


Figure 3a: linking a concept to a symbol

Figure 3b: definition by example of a new template

Figure 3c: browsing the language

The multimodal environment consists of the speech and sketch recognition modules. The sketch recognition module uses a set of graphic symbols, while the speech recognition module uses the parsing module to analyse the user's speech.

The first step in establishing the relationships between the multimodal user dialogue and the template set is the parsing of the user's speech (the parsing step in this environment is similar to the one seen in the configuration environment, except that the system parses the user's speech instead of a set of use cases).

The parser analysis constitutes the input of the template recognition module, which uses the set of templates defined in the configuration environment.

5. THE SYSTEM

The system was developed in Microsoft Windows XP environment, using Java 2 Platform Standard Ed. 5.0 as the programming language.

Figure 3 show the system's focal characteristics. As mentioned above, the language is defined by a set of templates containing a way in which a concept can be expressed.

Some key concepts can be expressed in both speech and sketch modes: the user can associate a word with a symbol (sketch mode) and a word expressed in the speech mode. This is demonstrated in Figure 3a: in this case, the user wants to relate the word "relation" and the symbol "distance" to the concept of relation.

Once the user has defined this relation s/he can define new templates as shown in Figure 3b: templates are defined "by example". That is, the user provides the system with the type of expression having a particular meaning through an example; the system picks the key concepts (in this example, the concept of relation) and lets the user express the rest by any words belonging to the same syntactic categories of the words in the example given. For instance, if s/he says, "we can define a weak relation" the template will be recognised.

Once the language has been defined, the user can browse the template set that makes up the language, as shown in Figure 3c.

Once the template set has been defined, the user can start multimodal communication with the system, thus speaking while drawing. After examining the text related to the user's speech and sketch, the system displays the matches it has found. Deictic expressions are solved by showing the object they point to.

6. CONCLUSIONS

This paper demonstrates the importance of natural language processing in the definition of multimodal languages. It addresses ambiguities by providing a tool for language definition. It shows that multimodal language is built from a template set that encapsulates a syntactic and conceptual structure, where concepts are expressed in the various modes and correspond to a precise semantic value.

The language is defined using a rule-oriented focus in constructing the templates, but is data-oriented in the definition of the overall language. This approach helps to reduce the problem of language ambiguity.

REFERENCES

[1] Oviatt, S.L. Mutual disambiguation of recognition errors in a multimodal architecture. In: Proceedings of CHI, pp. 576–583. ACM Press, New York (1999).

[2] Oviatt, S.L.: Taming recognition errors with a multimodal interface. CACM 43(9), 45–51 (2000).

[3] Oviatt, S.L. and Kuhn, K. Referential features and linguistic indirection, in multimodal language. In *Proceedings of the International Conference on Spoken Language Processing.* Sydney, ASSTA Inc., 2339–2342.

[4] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, Xiaoyang Mao: A Gaze and Speech Multimodal Interface. ICDCS Workshops 2004: 208-214.

[5] Henry Lieberman and Amy Chu, An Interface for Mutual Disambiguation of Recognition Errors in a Multimodal Navigational Assistant Multimedia Systems Journal, Special Issue on User-Centered Multimedia, Summer 2006.

[6] Sharon L. Oviatt, Antonella De Angeli, Karen Kuhn: Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. CHI 1997: 415-422.

[7] J. Bouchet, L. Nigay, T. Ganille: The ICARE Component-Based Approach for Multimodal Input Interaction: Application to Real-Time Military Aircraft Cockpits. HCI International, 3rd International Conference on Universal Access in Human-Computer Interaction, Las Vegas, Nevada, USA, Juillet 2005.

[8] Michael Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of COLING-ACL 1998*, pages 624–630.

[9] Roberto Navigli, Paola Velardi: **Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation, IEEE Transactions on Pattern Analysis and Machine Intelligence** Volume 27 , Issue 7 (July 2005) Pages: 1075 – 1086.

[10] WordNet 2.1: A lexical database for the English language; http://www.cogsci.princeton.edu/cgi-bin/webwn, 2005.

[11] Morris, J and Hirst, G (1991): Lexical cohesion computed by thesaural relations as an indicator of the structure of Text . *Computational Linguistics*, 17(1):21-42.

[12] Nigay, L.; Coutaz, J. 1993. A design space for multimodal interfaces: concurrent processing and data fusion. INTERCHI&#3993 Proceedings, Amsterdam, 172-178

[13] http://nlp.stanford.edu/software/lex-parser.shtml