Multimodal Sentence Similarity in Human-Computer Interaction Systems

Fernando Ferri, Patrizia Grifoni, and Stefano Paolozzi

Istituto di Ricerca sulla Popolazione e le Politiche Sociali - Consiglio Nazionale delle Ricerche, Via Nizza 128, 00198 Rome, Italy
{fernando.ferri,patrizia.grifoni,stefano.paolozzi}@irpps.cnr.it

Abstract. Human-to-human conversation remain such a significant part of our working activities because its naturalness. Multimodal interaction systems combine visual information with voice, gestures and other modalities to provide flexible and powerful dialogue approaches. The use of integrated multiple input modes enables users to benefit from the natural approach used in human communication. However natural interaction approaches may introduce interpretation problems. This paper proposes a new approach to match a multimodal sentence with a template stored in a knowledge base to interpret the multimodal sentence and define the multimodal templates similarity. We have assumed to map each multimodal sentence to the natural language one. The system then provides the exact/approximate interpretation according to the template similarity level.

Keywords: Human-Computer interaction, multimodality, sentence similarity.

1 Introduction

Face-to-face conversation remains such a significant part of our working activities despite of the availability of a great number of communication technologies. Anyway, there is a great interest towards natural interaction approaches and great efforts in development of technologies to aid such type of approaches (see for example [1], [2], [3]) and in improvement the interpretation on the computer side.

From this perspective multimodality has the potential to greatly improve Human-Computer Interaction combining harmoniously different communication methods. A Users can use voice, handwriting, sketching and gesture to input information. On the other side the system can use icons, text, sound and voice (output)to present information.

This paper uses the concept of multimodal language defined as a set of multimodal sentences [4], by the extension of the definition of Visual Language given in [5]. A multimodal sentence contains atomic elements (glyphs/graphemes, phonemes and so on) that form the Characteristic Structure (CS). The CS is given by the elements that form functional or perceptual units for the user. A multimodal sentence is defined, similarly to [6], as a function of: 1) the multimodal message, 2) the multimodal description that assigns the meaning to the sentence, and 3) the interpretation function

B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part II, LNAI 4693, pp. 403–410, 2007.
© Springer-Verlag Berlin Heidelberg 2007

that maps the message with the description, and the materialization function that maps the description with the message.

The goal of this paper is to propose a new approach to understand how a multimodal input sentence is matched with a template considering how different modalities cooperate each other, taking into account the user's behavior that can alter the multimodal input recognized by the system. The resulting multimodal input sentence is matched with a template stored in a knowledge base to provide an interpretation of the sentence. The sentence can precisely match the template or approximates it.

We consider the speech modality as the prevalent one because, generally, users explain their intentions by speech and use other modalities to "support" the speech and eventually to resolve ambiguities.

In our system each multimodal sentence corresponds to a natural language one. The system returns the interpretation of the multimodal sentence if the template that exactly maps with the corresponding natural language sentence is available. When the user interacts with the system, the corresponding multimodal sentence must refer to a stored template in the knowledge base in order to be interpreted. If the corresponding sentence, expressed in natural language, doesn't match with any of the stored templates than the system can interpret those sentences that approximate the matching considering templates similar to the first one. Because of some multimodal sentences (with different templates) can have very close interpretations (some times they can have the same meaning), this paper proposes to calculate the templates' similarity starting from the semantic similarity of the natural language sentences corresponding to the Multimodal one, we also take into account the user's behavior that can alter the multimodal input recognized by the system. For this purpose it is possible the association of a sentence and its template with the most similar template computing semantic similarity between natural language sentences. In this way we can provide an interpretation of the multimodal sentence also in case of non-perfect matching.

The natural language sentence can be represented as a semantic network of objects and binary relations among them, where each object corresponds to one node.

The paper is structured as follows: Section 2 provides a running example, section 3 describes the proposed approach. section 4 reports the conclusion.

2 Running Example

In order to explain our approach we consider the following example: the user draws an Entity-Relationship scheme assigning a label to each construct. Without loss of generality we assume that the input is given by only two modalities: speech and sketch. This scenario is represented in Fig. 1.

Suppose that the user sketches the diagram as shown in Fig. 1a, and he/she speeches the sentences shown in Fig. 1b. The system has to interpret the multimodal sentence and then has to materialize it as shown Fig. 1c.

For the sake of simplicity we only consider the creation of the Teaching Relation.

The user says: "The rhombus is the relation Teaching", at the same time the user sketches the figure of a rhombus (in his/her intention).

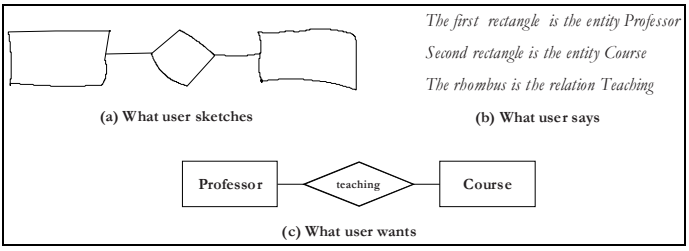

Fig. 1. An example of multimodal input

If the figure is properly interpreted by the sketch recognizer as a rhombus (Fig. 2a), we are in the typical situation of redundant information given by different modalities (i.e. the concept "rhombus" is expressed both by speech and sketch modalities) The Natural Language sentence associated to the multimodal one will be: "The rhombus is the relation Teaching" and it represents all the concepts expressed by speech and sketch.

Given such a sentence, a corresponding template to the sentence in the knowledge base, must be found.

The situation is more complex if we have some ambiguities problem. That is for example, if a user draws a figure that in his/her intention is a rhombus, recognized by the system as a different figure (Fig. 2b). So the system is unable to identify that the given input is redundant and can produce an incorrect multimodal sentence interpretation. In Fig. 2 both situations are illustrated with the proper timeline.

From the user point of view both situations must reproduce the same multimodal sentence, but due to ambiguities in sketch recognition the system does not produce the same sentences and they are not associated to the same template. In order to avoid these problems we propose an approach that take into account the user's behavior to reproduce multimodal sentences as near as possible to the user's input will.

3 Evaluating Multimodal Sentence Similarity

The first important problem to solve is to understand how different modalities cooperate and what is the template that the multimodal sentence matches according to the cooperation modality.

Six type of cooperation between modalities have been distinguished (see [7]): *Complementarity*, *Concurrency*, *Equivalence*, *Redundancy*, *Specialization* and *Transfer*.

Let us consider the interaction of two modalities (in particular we address speech and sketch modalities) M_1 and M_2 that transmit information in ΔT_1 and ΔT_2 time intervals respectively. The possible time intervals relationship between M_1 and M_2 are summarized in Fig. 3 and are:

- *Sequential*: The transmission of the second modality starts after the first one.
- *Disjoint*: The transmissions of the two modalities take place in two separated time intervals.

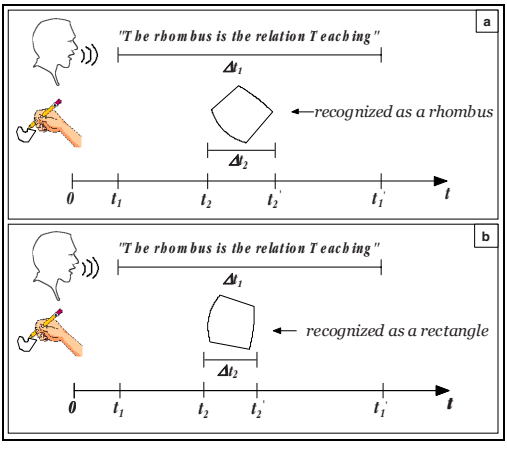

Fig. 2. Example of multimodal input (by speech and sketch) with timeline

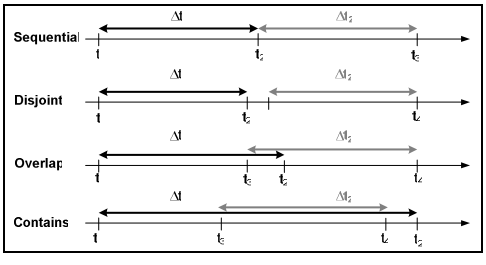

Fig. 3. Time intervals relationship in multimodal interaction

– *Overlap*: The transmission of a modality partially overlaps the transmission of the other one.
– *Contains*: The transmission of a modality is self contained in the transmission of the other one.

Evaluating time intervals for the involved transmission we analyze the possible combination of different input events. Multimodal input events can either be interpreted independently, or they can be merged.

Let us refer to the example given in the previous section. Firstly the system individually recognizes the concepts or the chunk of information for each modality involved in the input event. In this case the system must perform a speech and a sketch recognition in order to capture the initial information. Then each input modality is associated with its own time interval of transmission. The next step is to interpret these unimodal input on the base of the transmission time and the information recognized in order to extract a multimodal sentence representing the whole input event.

For our purposes, it is important to address two type of cooperation: Complementarity and Redundancy. Considering the aforementioned time intervals, Sequential, Overlap and Disjoint transmissions can denote Complementarity

interaction as Contains and Overlap transmissions can denote Redundancy interaction as previously stated in [7]. In Fig. 4 this relationships are presented.

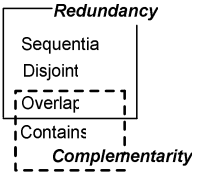

Fig. 4. Time intervals and type of interaction

In this paper we address redundancy, because it has been observed that a redundant multimodal input involving speech and sketch enables a more natural form of communication.

As stated before each modality transmission represents a finite number n of concepts.

Let us suppose that the modality M_1 transmits the concept C_1 in the time ΔT_1 and the modality M_2 transmits the concept C_2 in the time ΔT_2. ΔT_1 partially overlaps time ΔT_2. In our example M_1 is the speech modality and M_2 is the sketch modality. We have to identify the template for the multimodal sentence in order to correctly interpret it. That is, this "match" requires interaction between modalities has to be recognized. In redundant interaction C_1 and C_2 must be the same concept. However, the modality used to express one concept (for example concept C_1) can introduce some imprecision and approximations. This is the situation in which the user would draw a rhombus, but the system recognizes a rectangle due to the drawing imprecision. At the same time the concept expressed by speech mode is "the rhombus". In this case the two concepts C_1 and C_2 are similar concepts. The question is: what similarity measure has to be considered between two concepts in order to have a redundant interaction between modalities?

This measure car vary among different users. The system has to acquire knowledge on the user's behavior during the whole interaction process.

We have considered a sample of 20 different users, and we asked them to draw a rhombus, alternatively with other figures, using our sketch interface, for 20 times. This has permitted us to take into account the user's behavior, that is used to better calculate the similarity between concepts.

The sketch recognizer identifies, for each sample, a measure of similarity between user's sketch and the required figure (in this case a rhombus).

This measure represents a fundamental element for the computation of concept similarity. More formally, if C_i is the *i*-th sample drawn by the user, and n is the total number of the samples, the user behaviour approximation A_{ss} (for sketch-speech modalities interaction) is:

$$A_{SS} = \frac{1}{n}\sum_{i=1}^{n} C_i \,, \qquad A_{ss} \in (0,1) \,. \tag{1}$$

The concepts are stored in the concepts database as natural language terms. We adopt an extended words similarity algorithm in order to calculate semantic similarity between concepts using the WordNet lexical database [8].

3.1 Evaluating Concepts Similarity

The taxonomical structure of the WordNet knowledge base is important in determining the semantic distance between words. In WordNet, terms are organized into synonym sets (synsets), with semantics and relation pointers to other synsets.

One direct method for similarity computation is to find the minimum length of path connecting the two words [9].

However, this method may be not sufficiently accurate if it is applied to a large and general semantic net such as WordNet. To address this weakness, it is important to notice that concepts at upper layers of the WordNet's hierarchy have more general semantics and less similarity between them, while words that appear at lower layers have more concrete semantics and have a higher similarity. Therefore, also the depth of word in the hierarchy should be considered. In summary, we note that similarity between words is determined not only by path lengths but also by depth (level in the hierarchy). Moreover we take into account the user behaviour considering the user behaviour approximation Ass. The proposed algorithm is an extensions of the one proposed in [10] for words similarity.

Given two concepts, c1 and c2, we need to find the semantic similarity s(c1,c2).

Let be l the shortest path length between c1 and c2, and h the depth of subsumer in the hierarchical semantic nets, the semantic similarity can be written as:

$$s(w_1, w_2) = f_1(l) \cdot f_2(h) \cdot B_{ss} \,. \tag{2}$$

B_{ss} represents the user's behaviour contribution and its value depends on the value of the product between $f_1(l)$ and $f_2(h)$. If this value is higher than a threshold value (δ), B_{ss} is equal to $(A_{ss})^{-1}$, otherwise B_{ss} is equal to 0. More formally:

$$B_{ss} = \begin{cases} 0 & \text{if} \quad f_1(l) \cdot f_2(h) < \delta \\ \dfrac{1}{A_{SS}} & \text{if} \quad f_1(l) \cdot f_2(h) \geq \delta \end{cases} . \tag{3}$$

The path length between two concepts, c_1 and c_2, can be computed according to one of the following cases:

1. c_1 and c_2 belong to the same synset,
2. c_1 and c_2 do not belong to the same synset, but their synsets contains one or more common words,
3. c_1 and c_2 neither belong to the same synset nor their synsets contain any common word.

By the above considerations we can considered the function $f_1(l)$ to be a monotonically decreasing function of path l.

For depth contribution, function $f_2(h)$ should be a monotonically increasing function with respect to depth h.

The whole formula for words similarity computation is:

$$s(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \cdot B_{ss} \,. \tag{4}$$

The values of α and β depend on the used knowledge base. For WordNet the optimal parameters are α=0.2 and β=0.45 as explained in [11].

3.2 Evaluating Multimodal Sentence Similarity

The interpretation of a multimodal sentence needs its corresponding sentence, expressed in natural language form, matches with one template contained in the knowledge base. An approximate interpretation, which uses a semantic approach, can be provided.

As discussed in the introduction, the need of interpreting sentences characterizing the multimodal dialog has led us to propose a process, devoted to obtain the exact interpretation or its approximation. The steps of our algorithm for evaluating multimodal sentence similarity can be summarized as follows:

- the user formulates his/her multimodal sentence,
- the resulting sentence is transformed in a Natural Language (NL) one and its template is compared with the templates sharing the same keywords (this reduce the number of the compared templates); the corresponding templates are selected,
- if there are no matches, we extract from each one of the selected templates the relative sentences that have been used to create (by example) it,
- then the system computes the semantic similarity between each of these sentences and the user's sentence. Indeed sentence similarity is calculated through computing similarity between templates of each sentence. After investigating a number of methods, we proposed a templates similarity measure following the approach reported in [10].
- the highest value of semantic similarity is used to choose the more similar template with the given sentence.

4 Conclusions

Multimodal human-computer interaction, in which the computer accepts input from multiple channels or modalities, is more flexible, natural, and powerful than unimodal interaction with input from a single modality. However naturalness of communication is directly proportional with the complexity of the interpretation of messages. In this paper is presented an approach to define how a multimodal sentence matches to a sentence's template according to: 1) the different type of cooperation between modalities; 2) the similarity approximations given by the user's behaviour in his/her multimodal input. We have implemented a multimodal system based on speech and sketch modalities that uses the aforementioned methodology to better understand user's input, also considering the user's way of inputting. Future work will address other input modalities, such as gesture, that could also help disambiguate the user's interaction behaviour.

References

1. Binot, J.L., Falzon, P., Perez, R., Peroche, B., Sheehy, N., Rouault, J., Wilson, M.D.: Architecture of a multimodal dialogue interface for knowledge-based systems. In: Proceedings of Esprit'90 Conference, pp. 412–433. Kluwer Academic Publishers, Dordrecht (1990)
2. Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A.: Unification-based Multimodal Integration. In: Proceedings of the ACL, Madrid (1997)
3. Wahlster, W.: User and discourse models for multimodal communication. In: Sullivan, J., Tyler, S. (eds.). Intelligent User Interfaces. ACM Press, Addison-Wesley (1991)
4. Caschera, M.C., Ferri, F., Grifoni, P.: Multimodal interaction systems: information and time features. Journal of Web and Grid Services (to appear)
5. Bottoni, P., Costabile, M.F., Levialdi, S., Mussio, P.: Formalizing visual languages. VL 1995, 45–52 (1995)
6. Celentano, A., Fogli, D., Mussio, P., Pittarello, F.: Model-based Specification of Virtual Interaction Environments. In: Proceedings of the 2004 IEEE Symposium on Visual Languages - Human Centric Computing (VLHCC'04) - vol. 00, pp. 257–260 (2004)
7. Martin, J.C.: Toward intelligent cooperation between modalities: the example of a system enabling multimodal interaction with a map. In: Proceedings of Int. Conf. on Artificial Intelligence (IJCAI'97) Workshop on Intelligent Multimodal Systems, Nagoya, Japan (1997)
8. WordNet 2.1: A lexical database for the English language (2005), http://www.cogsci.princeton.edu/cgi-bin/webwn
9. Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. IEEE Trans. System, Man, and Cybernetics 9(1), 17–30 (1989)
10. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on knowledge and data Engineering 18(8), 1138–1150 (2006)
11. Li, Y.H., Bandar, Z., McLean, D.: An Approach for Measuring Semantic Similarity Using Multiple Information Sources. IEEE Trans. Knowledge and Data Eng. 15(4), 871–882 (2003)