International Journal of Computational Intelligence Systems
Vol. **13(1)**, 2020, *pp.* 178–192
DOI: https://doi.org/10.2991/ijcis.d.200208.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

DAMA: A Dynamic Classification of Multimodal Ambiguities

Patrizia Grifoni, Maria Chiara Caschera*, Fernando Ferri

National Research Council, Institute of Research on Population and Social Policies (CNR-IRPPS), Via Palestro 32, 00185, Rome, Italy

ARTICLE INFO

Article History
Received 13 Mar 2019
Accepted 07 Feb 2020

Keywords
Hidden Markov models
Human–machine interaction
Multimodal interaction
Natural language processing

ABSTRACT

Ambiguities represent uncertainty but also a fundamental item of discussion for who is interested in the interpretation of languages and it is actually functional for communicative purposes both in human–human communication and in human–machine interaction. This paper faces the need to address ambiguity issues in human–machine interaction. It deals with the identification of the meaningful features of multimodal ambiguities and proposes a dynamic classification method that characterizes them by learning, and progressively adapting with the evolution of the interaction language, by refining the existing classes, or by identifying new ones. A new class of ambiguities can be added by identifying and validating the meaningful features that characterize and distinguish it compared to the existing ones. The experimental results demonstrate improvement in the classification rate over considering new ambiguity classes.

1. INTRODUCTION

In human–human interaction, the multiple interpretations of messages consisting of words, gestures, gazes and, generally input of other combined modalities, offer to the imagination a wide variety of topics for reflection and interpretation. However, multiple interpretations produce ambiguities. At the same time, ambiguities represent uncertainty but also a fundamental item of discussion for philosophers, lexicographers, linguists, cognitive scientists, literary theorists and critics, authors, poets, orators, computer scientists and for everyone who is interested in the interpretations of languages. Piantadosi *et al.* [1] underlined that ambiguity is actually functional for communicative purposes, rather than dysfunctional, and "any efficient communication system will necessarily be ambiguous." Similarly, Negroponte considers ambiguity at the basis of approaches to overcome the limitation of the "sensory deprived and physically limited" interaction systems. But ambiguities also imply the need to be identified, managed and, then, solved. Identifying a specific class of ambiguities is an important step, as it means detecting the features to be managed, facilitating the ambiguity solution and, therefore, supporting the interpretation process. Indeed, when classes of ambiguities are detected, information characterizing them can be computed simplifying the solution process.

The concept of ambiguity has been widely discussed in the literature both from the point of view of human–human communication [2–4] and of human–machine interaction [5–8]. Many studies on Natural Language Processing, studies on human–machine interaction and multimodal systems (i.e., systems that use "two or more combined user input modes — e.g., speech, pen, touch, manual gesture, gaze and head and body movements — in a coordinated manner with multimedia system output" [9]) aim to replicate powerfulness, flexibility and naturalness of human–human communication process. The paradigm shift to multimodal systems reflects the evolution of more expressively powerful computer input that adapts to human cognition and performance. Discussions on ambiguities in multimodal systems, provided in [10–12], proposed static rule-based approaches for classifying multimodal ambiguities. However, as the interaction language evolves [13], the ambiguities can evolve, and the need to refine the identification of the existing classes, but also to define a new one, can appear. This requires a dynamic classification of multimodal ambiguities.

The main contributions of this paper can be summarized as follows:

- The method is proposed to dynamically classify multimodal ambiguities representing them as sequences composed of several modal information using a linguistic approach;
- The method classifies the multimodal ambiguities by modelling each ambiguity as a Hidden Markov Model (HMM). The proposed method is able to capture the evolving nature of the interaction language by learning the evolution of the interaction language, refining the existing classes, or identifying new ones. The experimental results demonstrate improvement in the classification rate over considering new ambiguity classes.

The current paper is structured as follows. Section 2 describes the problem statement that has driven this work. Section 3 gives an overview of the literature works on classification methods. Section 4 presents some preliminary notions needed to describe the principles of the method defined in this paper. Section 5 presents the addressed problem of multimodal ambiguity classification. Section 6 describes the multimodal ambiguities classifier.

*Corresponding author. Email: mc.caschera@irpps.cnr.it

Section 7 shows the results of the evaluation test of the accuracy of the proposed method. Finally, Section 8 concludes the paper and discusses future works.

2. PROBLEM STATEMENT

An ambiguity results when the same structural form of a language has more than one meaning [14]. "The structural form contains the word, the phrase, the sentence, the discourse and the utterance, while the meaning not only refers to conceptual meaning, connotative meaning, social meaning, affective meaning, reflected meaning, collocative meaning and thematic meaning, but also the meaning in use" [15]. In addition, considering the complexity of the multimodal human language but also the humans' ability to be ironic, sarcastic or lie (for citing some examples), we understand how human behavior and human language can be complex and ambiguous to be interpreted by other humans or by a multimodal interaction system. The human language is inherently ambiguous and it implies the need to acquire knowledge about the ambiguity before the recognition process of human–machine interaction. Indeed, the ambiguity issue turns out to be relevant in different application domains such as e.g., security and surveillance, but also human–machine interaction, assistance systems, or infotainment systems for in-car interaction. As an example, if the selection of an object (e.g., window) or a function in the car environment is ambiguous, the system needs to ask driver until the object and function selection is well-defined. In addition, the human language evolves [16]. The capacity of the language to evolve implies changes that were initially caused by cultural changes and were subsequently defined through the evolutionary capacity to adaptively respond to new communicative conventions [17]. Cultural changes and new communicative conventions mainly influence ambiguities connected on how the meaning of a sentence depends on its function in everyday life, i.e., the larger context of the conversation. These changes mainly affect ambiguities connected to the user intention, sentiment, pragmatic knowledge (i.e., how sentences are used in different situations and how use affects the interpretation of the sentence), discourse knowledge (i.e., how preceding sentences identify the meaning of a sentence) and world knowledge (i.e., users' beliefs and goals in a conversation). The pragmatic ambiguity belongs to this type of ambiguity because it "refers to a situation where the context of a sentence gives it multiple interpretations" [18]. An example of pragmatic ambiguity is given with the following sentence uttered by two people; Jenny and John that are talking in the kitchen during the dinner [19]. Jenny says:

``Put this on the plate and eat it''

Jenny indicates the carrot on the plate, which already contains other vegetables, while she is saying:

``This.''

"It" can refer both to the carrot and the plate, which is intended as all the vegetables it contains, and the sentence has two meanings in the context in which it is uttered [18]. This is a pragmatic ambiguity; as stated in [20] a "pragmatic ambiguity arises when the statement is not specific, and the context does not provide the information needed to clarify the statement." The process of ambiguity management implies the need to model the knowledge about the ambiguity to acquire the features of evolving of the interaction languages.

This paper deals with the identification of the meaningful features of multimodal ambiguities classes and their dynamic identification. The role of ambiguity classification process for human–machine interaction has been explained in [10] as the first step of a systemic approach to the solution process of any ambiguous multimodal sentence (i.e., the linguistic unit structuring the different multimodal inputs) [21]. The introduction of a classificatory step before the ambiguity solution allows adopting a systematic and modular approach. We start from the idea that an incorrect (i.e., ambiguous) interpretation implies the identification of the meaningful features to be managed for solving the ambiguity [22]. This paper goes beyond the static classification process proposed in [10] and provides a dynamic approach modeling knowledge about multimodal ambiguities. It keeps up with the evolution of language and the ability to refine the identification of existing classes but also defining new ones.

Ambiguity classes may evolve considering that: i) features, which characterize classes, may change; and ii) new classes may be created when ambiguities, which are produced by the language evolution, are intercepted and they cannot be included in the existing classes. That is, features and constraints can evolve as: i) new constraints can be defined to match an ambiguous sentence in an existing class, and ii) new features and constraints can be defined to identify new classes of ambiguities. In particular, considering multimodal interaction, a multimodal ambiguity may be inherited from a modal ambiguity when an incorrect interpretation of an input belonging to a modality affects the correct interpretation of the multimodal input that contains the modal input with an incorrect interpretation (propagation from modal to multimodal level) [23]. An example of multimodal inherited from a modal ambiguity is given by a user that says by speech:

``Show this near school''

While she/he's selecting both the icon of a hotel and the icon of a restaurant by sketch (modal ambiguity) [10]. In this case, the multimodal ambiguity is generated by the modal (sketch) ambiguity that is due to the inability to establish if the user is selecting the icon of a hotel or the icon of a restaurant while she/he is uttering "this." In addition, the multimodal ambiguity may be a consequence of the combination of the modal inputs that are correctly recognized at modal level but, that is not coherent at multimodal level. In fact, information coming from each separate modality in input can be correctly and univocally interpreted by the multimodal system, while the interpretation can become ambiguous by considering combined information [12]. An example of this kind of ambiguity is represented by a user that says by speech.

``This is a river''

While she/he's drawing a lake by sketch; in this case, the combination of river and lake generates incoherence between the concepts connected with the two modal inputs, producing a multimodal ambiguity [10].

For this reason, a multimodal input inherits a modal ambiguity but it can be also affected by new ones. A detailed discussion on multimodal ambiguities is provided in [10,11]; this paper enriches the debate investigating how to address the potential evolution of the ambiguities produced by the evolution of the way to interact and communicate as well as of interaction languages [13]. Motivated by the need to automatically detecting new ambiguity classes,

this paper investigates and proposes a method for modeling the dynamic nature of the interaction process.

3. RELATED WORKS

Since managing ambiguity is a complex process, the identification of a specific type of ambiguities enables supporting the interpretation process by detecting the features to be managed and, therefore, optimizing the solution process [21]. In particular, the management of ambiguities can be divided into two sub problems: the first one deals with the identification of the features to be managed through the ambiguity classification step, and once identified, these features may be appropriately managed through the ambiguity solution step (second sub problem). Dividing the problem into sub problems allows managing more efficiently complex processes. Several works were performed on ambiguity detection and solution process [18,24]. Some of the works aimed at identifying typical ambiguous terms and constructions of terms [25,18]; while other works were focused on the solution process using the natural language understanding methodologies [26] and artificial intelligence and statistical techniques [12,27]. Note that several studies focused on the definition of the ambiguity classes for Visual Languages [6,8], other for Natural Language [5] and finally for Multimodal Language [10]. Some studies were related to ambiguities in a legal text [28] and in the context of visual surveillance [29].

Since this paper focuses on the classification process; the purpose of this section is providing an overview of the most relevant classification methods. In this direction, several research efforts have been undertaken by machine learning and statistics communities [30]; they usually aim at providing classification methods for organizing a collection of objects. A classification method can serve as an explanatory tool both to distinguish among objects of different classes and to predict the value (or the class) of a user specified goal attribute, based on the values of predicting attributes [31].

In detail, classification methods are built from an input data set by classification techniques that can be divided in [32]: i) methods based on generation of models with separate model components, each one explaining part of a given dataset, e.g., decision tree induction [10,33], the lattice machine [34] and rule-based classifiers [35]; ii) and, paradigms that do not build models with separate parts, i.e., Bayesian Network (BN) [36], Support Vector Machines (SVMs) [37], Neural Networks (NNs) [38] and HMMs [39].

The decision tree is a natural and intuitive paradigm that allows classifying patterns through a sequence of questions. In this paradigm, trees classify instances by sorting them based on feature values; it is a widely used practical method based on inductive inference [40]. This paradigm is usually used for classification because the leaves of the tree represent classifications and the branches represent feature-based splits that lead to the classifications.

Similarly, in lattice machine-based method, data are structured by relations and are represented by a subset of the elements in the lattice through a partition of the datasets into classes [38].

Rule induction is a special kind of machine learning technique reasoning from specific to general principles (expressible as if-then rules) [41,42].

Differently from the decision trees, BNs [36] represent the structural relationships among their features taking into account prior information about a given problem (e.g., if a node is direct cause or effect of another node or if two nodes are independent). Indeed, BNs are represented by direct acyclic graphs with one parent and several children associated with a set of variables, the features. Among child nodes and their parents, there are probability relationships; arcs represent casual influences. The lack of arcs indicates conditionally independencies. A disadvantage of BNs is that they are not suitable for datasets with many features [43] because of the complexity for a very large network in terms of time and space. For example, in [44], BNs are used to classify ambiguities from face, body and speech data; a separate Bayesian classifier is used for each modality, i.e., face, gesture and speech.

Differently from BNs, SVMs [37] are well suitable to deal with a large number of features of the training dataset, because the SVMs is unaffected by this number of features. However, SVMs need a large sample size to achieve their maximum prediction accuracy [45].

Similarly, to SVMs, NNs [38] tend to perform much better when dealing with multi dimensions and continuous features [45]. The advantage of NNs consists in the fact that they can approximate any function with arbitrary accuracy as they are able to adapt themselves to the data without any explicit specification of functional or distributional form for the underlying model and they are flexible in modeling complex relationships because they are a nonlinear model. As example, in [46] a NN-based approach was applied to merge and combine decisions made disjointly by the audio unit and the visual unit of a multimodal system.

HMMs [39] allow modeling the dynamics (i.e., the process) of the modeled issue, and to efficiently estimate parameters by maximizing the likelihood of data given the model. HMMs have been widely used for temporal pattern recognition, i.e., for classifying speech recognition [47], gesture classification [48] and for semantic analysis in the case of handwriting recognition [49], part-of-speech detection [39], audio-visual speech recognition [50] and classification of different types of videos [51].

Comparing approaches, SVMs [37] and NNs [38] can achieve a good accuracy on classification tasks; however, if the set of classes' changes, generative probabilistic models for learning classes' conditional distribution are more appropriate. Among generative probabilistic models, we decided to use HMMs [39] as they allow modeling sequences of structured data.

By using HMMs, we are able to represent differences in the whole structure of multimodal sentences. They have been applied to model and classify dialogue pattern [52–55], and to classify multimodal events adopting multimodal features and incorporating temporal frequent pattern analysis for baseball event classification [56].

Therefore, we use HMMs because they enable dynamically modelling complex data and, therefore, they are well suited for our purposes: i) to build a method aiming to automatically classify any ambiguity; ii) to identify any new multimodal ambiguity class and, therefore, iii) to progressively learn the evolution features of the language. In fact, HMMs enable dynamically modelling of the multimodal ambiguities classification process.

The description of DAMA (a Dynamic clAssification of Multimodal Ambiguities) needs of some background fundamental notions provided in the following section.

4. FEATURES EXTRACTION

Since the proposed method is based on the representation of each multimodal sentence as a string of symbols of the used language, some concepts and notions need to be introduced. Fundamental is the concept of *multimodal sentence* [21], which is composed of a set of *terminal elements* that are the elementary parts of a multimodal language [57]. As defined in [21], each *terminal element* (E^i) is identified by a set of distinctive attributes that we refer as features. These features are: E^i_{mod} that identifies the modality used by the element (meaning the speech, sketch, etc.); E^i_{repr} i.e., the representation of the element; E^i_{time} i.e., the temporal interval (given by the start and end time values) connected to the element E^i; E^i_{role} that specifies the syntactic role that the element E^i plays in the multimodal sentence according to the Penn Treebank Tag set [58] (e.g., a noun, a verb, an adjective, an adverb, a pronoun, a preposition, etc.) and $E^i_{concept}$ that specifies the concept associated with the element, providing its semantics (i.e., the application domain).

Any information about the interaction modalities, the temporal intervals and the elements representations are extracted during the user's interaction with the system. This information is extracted by technologies for the gesture, facial expression, speech and handwriting recognition (see Figure 1). On the other hand, information about the concepts connected to the input is given using an ontology defined according to the interaction context. The explanation of how this information is extracted is beyond the scope of this paper.

To clarify the introduced concepts, we consider the example cited in Section 2, in which two people, Jenny and John, are talking in the kitchen during the dinner. Jenny says to John:

``Put this on the plate and eat it''

While she is indicating the plate, which contains the carrot and other vegetables.

Figure 2 shows the sentence with its timeline.

All the elements defined through the interaction modalities (in the example speech and gesture) are combined in the multimodal sentence. The verbal (speech) and not verbal (gesture) elements are combined when they are synchronized.

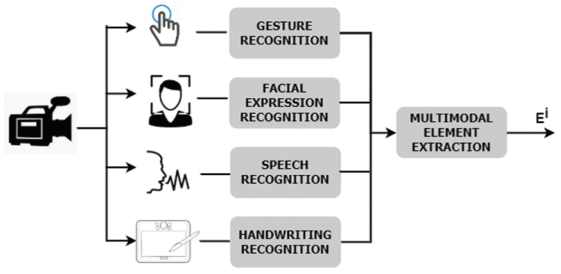

Figure 1 | Extraction process of terminal elements.

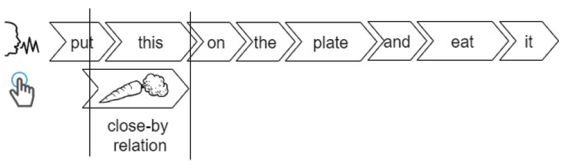

Figure 2 | Example of the ambiguous multimodal sentence.

As described in [59], synchronized elements can be identified comparing the temporal intervals when they are created (i.e., E^i_{time}) and combining them when they are in CloseBy relation [23] (i.e., by computing temporal differences between the beginning and ending of each interval).

Figure 3 shows the terminal elements that compose the multimodal sentence in Figure 2 and their features (i.e., E^i_{mod}, E^i_{repr}, E^i_{time},E^i_{role}, $E^i_{concept}$).

Note that, the resulting multimodal sentence consists of eight elements by the speech, one by the gesture and one null element as Figure 3 shows. The syntactic roles and the syntactic dependencies between the elements of a multimodal sentence are extracted using the Stanford Parser (nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/parser/lexparser/LexicalizedParser.html), which parses the natural language sentence associated with the multimodal sentence by a linearization process [21]. The knowledge about the syntactic roles and the syntactic dependencies allows building the syntax-tree that represents the syntactic structure of the sentence. When a multimodal sentence is ambiguous, it can be associated with more than one syntax-tree, one for each interpretation of the sentence in Natural Language.

All the syntax-trees are combined in a direct acyclic graph to which we refer, hereafter, by the term *syntax-graph* [21]. This graph collapses common structures of the different syntax-trees associated

	E^i_{mod}	E^i_{repr}	E^i_{time}	E^i_{role}	$E^i_{concept}$
E^1	speech	put	(1,2)	vb1	put
E^2	speech	this	(3,5)	dt1	carrot
E^3	gesture		(2,5)	nn1	carrot
E^4	speech	on	(5,6)	in1	on
E^5	speech	the	(7,8)	dt2	the
E^6	speech	plate	(9,10)	nn2	plate
E^7	speech	and	(11,12)	cc1	and
E^8	speech	eat	(12,14)	vb2	eat
$E^{9'}$	speech	it	(14,15)	prp1	carrot
$E^{9''}$	speech	it	(14,15)	prp1	plate
$E^{10'}$	null	null	null	nn1	carrot
$E^{10''}$	null	null	null	nn2	plate

Figure 3 | Elements of the multimodal sentence in Figure 2.

with the multimodal sentence. In addition, each terminal node of the syntax-graph is a terminal element of the multimodal language, and each terminal node includes information about the specific element (i.e., E^i_{mod}, E^i_{repr}, $E^i_{concept}$).

In order to introduce the concept of syntax-graph, the formalism, which was defined in [21], is here mentioned.

Definition 1. *A syntax-graph is a direct acyclic graph that combines all the syntax-trees of the multimodal sentence, collapsing each common sub-tree of the different trees into one sub-tree only, and adding all noncommon nodes and arcs belonging to trees. It has the terminal elements of the grammar as terminal nodes.*

Figure 4 shows the syntax-graph of the multimodal sentence in Figure 2. According to the definitions of complementarity [23], the speech element "  this" and the gesture element "  " (carrot) are complementary, while the speech element "it" (E^9) is a pronoun that does not clearly specify the object (in this case can be referred to the carrot ($E^{9'}$) or to the plate($E^{9''}$)) of the action "eat." Therefore, in the corresponding syntax-graph, there is an element (E^{10}) that may be referred either to the carrot ($E^{10'}$) or to the plate ($E^{10''}$), but that is not defined by modalities (speech or gesture in the example). In this case, the features, associated to the modality and the representation, assume the value null. Note that the multimodal sentence of Figure 2 is ambiguous because the pronoun corresponding to the speech element "  it" can be referred both, to the concept carrot or plate (see Figure 4). This is an example of pragmatic ambiguity and, in particular, a co-reference ambiguity. This pragmatic ambiguity arises because the statement "eat it" is not specific, and the information about the object referred to the pronoun "it" by the action "eat" is missing, and must be inferred. In order to have an unambiguous interpretation of the multimodal sentence, the pronoun "it" should refer to one concept only. In this case, however, the pronoun "it" refers to two different concepts (i.e., carrot and plate) and the syntax-graph (see Figure 4) has two elements that have the same syntactic role, but refer to two different concepts. The syntax-graph highlights that the multimodal sentence has two possible interpretations (i.e., "put this carrot on the plate and eat the carrot" and "put this carrot on the plate and eat the plate").

In summary, each multimodal sentence is modeled as a sequence of m elements ordered according to the linearization process described in [56]. Each element (E^i) is characterized by a set of features that are: its modality (E^i_{mod}), its representation (E^i_{repr}), its temporal information (E^i_{time}), its syntactic role (E^i_{role}) and its concept ($E^i_{concept}$).

For our purpose, we consider the sequence of features that characterize the elements, which compose the ambiguous sentence, as the features vector f_t at time t related to the multimodal sentence compose of m elements. In detail,

$$f_t = [(E^v_r)] \quad \text{with } v \in M \text{ and } M = \{1..m\} \quad r \in R \text{ and } R = \{mod, repr, time, role, concept\} \tag{1}$$

In the example of Figure 2, the features vector f_t of the multimodal sentence consists of a sequence of m = 10 elements as Figure 3 shows.

The concepts and notions, introduced in this section, are used in the following sections to describe the classification method for multimodal ambiguities.

5. THE CLASSIFICATION PROBLEM

Before discussing the classification problem, it may be useful to clarify its role in the interpretation process of the multimodal input. The multimodal inputs are processed by unimodal recognition modules (speech, handwriting, sketch, etc.), and the recognized signals for the various modalities are integrated and interpreted according to a multimodal grammar [57]. The interpretation of the multimodal input could not be unique (i.e., it produces more than one interpretation of the user's input), than an ambiguity needs to be faced. If the interpretation process produces more than one interpretation of the user's input, then the classification process manages the ambiguous input interpretation providing its classification.

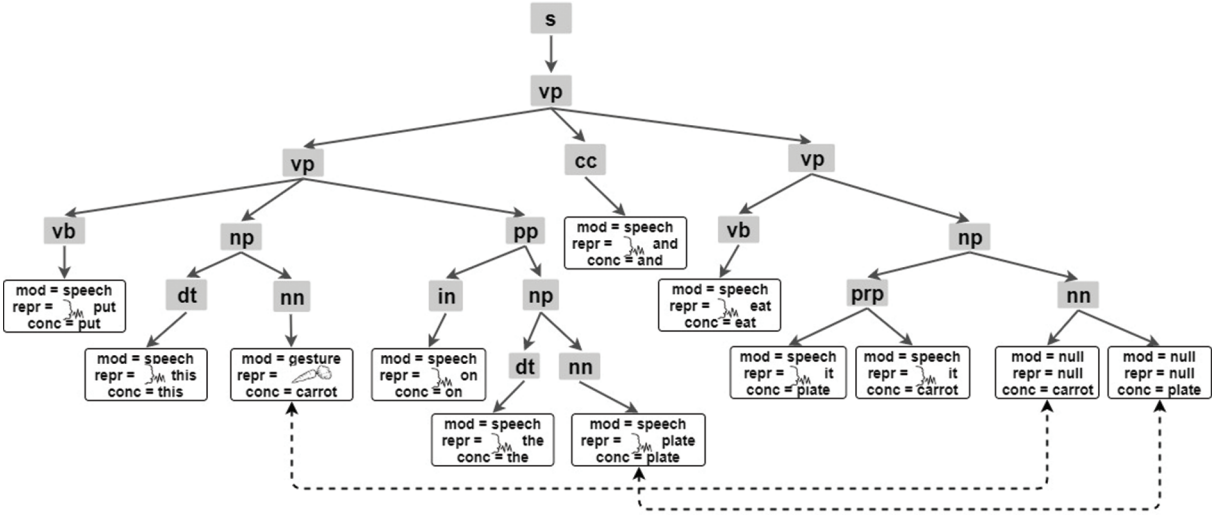

Figure 4 Syntax-graph of the multimodal sentence in Figure 2.

Then, information about the classified ambiguous input and the set of candidate interpretations are used to solve ambiguities [21], which selects only one interpretation, and, therefore, the ambiguity is solved.

In this paper, we address the classification step of ambiguities, and, therefore, we start from the assumption that only ambiguous sentences should be processed by the framework proposed in this paper. To allow understanding the method described in this paper, it is useful to briefly mention the different classes of multimodal ambiguities that have been widely discussed in [10], since we start our dissertation on the basis of the classification of ambiguities that authors proposed in [10].

Ambiguities can arise at the syntactic or at the semantic level.

The syntactic ambiguities are connected with the structure of the multimodal sentence. They appear when alternative syntactic structures (syntax-trees) for the multimodal sentence are generated during the interpretation process. In particular, these ambiguities occur when the role, which an element of the sentence plays during the interaction, is not univocally defined, and the elements of a multimodal sentence can be syntactically combined in more than one way. The syntactic ambiguities are classified as:

- *Gap ambiguity:* arising when an element of the multimodal sentence is omitted; therefore, when there is a terminal node in the syntax-graph that corresponds to a terminal element that has some features instantiated with the value null.
- *Analytic ambiguity:* arising when the syntactic categorization of the element is itself not univocally defined; therefore, when there are two different edges in the syntax-graph that can reach the same element in the syntax-graph.
- *Attachment ambiguity:* arising when a subset of the elements belonging to the sentence can be syntactically attached to two different parts of the sentence; therefore, when there are two different syntactic paths in the syntax-graph that can reach the same sub-tree in the syntax-graph.

The *semantic ambiguities* involve the meaning of the whole sentence or of a single element, and they are classified as follows:

- *Lexical ambiguity:* arising when one element has more than one generally accepted meaning; therefore, when there are two elements in the syntax-graph that have the same syntactic role, the same representation defined in the same modality, but they refer to different concepts;
- *Temporal-semantic ambiguity:* arising when two different elements of a multimodal sentence have the same syntactic role but they refer to two different concepts through different modalities. Since the other ambiguity classes are based on established definitions of ambiguities in Natural Language and Visual Language, this type of ambiguity needs further clarification. This ambiguity is produced by a combination of the modal inputs that are correctly recognized (not ambiguous) at modal level but, the interpretation of the combined modalities, which are in CloseBy relation, is not coherent (ambiguous) at multimodal level. This means that information coming from each separate modality in input can be correctly and univocally interpreted, while the interpretation becomes ambiguous by considering combined information [12]. In this case, there are two elements in the syntax-graph that have the same syntactic role, two different representations defined by two different modalities, and they refer to different concepts.
- *Target ambiguity:* arising when the user's focus is not clear; therefore, when there are two elements in the syntax-graph that have the same syntactic role, have two different representations defined in the same modality, but refer to different concepts.

This classification allows detecting the meaningful features of the multimodal ambiguous sentence that characterize the ambiguity class. In particular, it appears that it is possible to distinguish the ambiguities classes by introducing variables that allow distinguishing the introduced classes. The identified distinctive variables are:

- N: is a variable that identifies if there is an element (E^v) that has some features (E^v_r) that are not specified and that assume the value null ($E^v_r = \text{null}$). In this case the variable assumes the value N_1, otherwise N_0 (see the example in Figure 3).
- CBR: is a variable that identifies if two elements ($E^v \neq E^w$) of the sentence are in the CloseBy relation (E^v_{time} and E^w_{time} are CloseBy as defined in [23]). In this case, the variable assumes the value CBR_1, otherwise CBR_0.
- ESG: is a variable that identifies if there are two different edges ($E^v_{\text{role}'} \neq E^v_{\text{role}''}$) in the syntax-graph that can reach the same element (E^v) in the syntax-graph. In this case the variable assumes the value EGS_1, otherwise EGS_0.
- PSG: is a variable that identifies if there are two different syntactic paths in the syntax-graph that can reach the same sub-tree (a set of different elements $E^v \ldots E^w$ with $E^v_{\text{role}'} \neq E^v_{\text{role}''} \ldots E^w_{\text{role}'} \neq E^w_{\text{role}''}$) in the syntax-graph. In this case the variable assumes the value PSG_1, otherwise PSG_0.
- SR: is a variable that identifies if there are two elements ($E^v \neq E^w$) in the syntax-graph that have the same syntactic role ($E^v_{\text{role}} \equiv E^w_{\text{role}}$). In this case the variable assumes the value SR_1, otherwise SR_0.
- C: is a variable that identifies if there are two elements ($E^v \neq E^w$) in the syntax-graph that refer to same concepts ($E^v_{\text{conc}} \equiv E^w_{\text{conc}}$). In this case the variable assumes the value C_1, otherwise C_0.
- R: is a variable that identifies if there are two elements ($E^v \neq E^w$) in the syntax-graph that have the same representation ($E^v_{\text{repr}} \equiv E^w_{\text{repr}}$). In this case the variable assumes the value R_1, otherwise R_0.
- M: is a variable that identifies if there are two elements ($E^v \neq E^w$) in the syntax-graph that are defined in the same modality ($E^v_{\text{mod}} \equiv E^w_{\text{mod}}$). In this case the variable assumes the value M_1, otherwise M_0.

The following section presents how the described variables allow distinguishing the different ambiguity classes by capturing differences and modeling the variations among them.

6. DAMA: A DYNAMIC CLASSIFICATION OF MULTIMODAL AMBIGUITIES

The purpose of this paper is to address the classification step of ambiguities, and, therefore, we start from the assumption that only ambiguous sentences should be processed by the method proposed in this paper. Therefore, the paper addresses the needs: to intercept ambiguities by identifying their features, and to address the evolution of the interaction language. To address those needs, this paper presents a method that is able to identify and define the different classes of multimodal ambiguity in a dynamic way and, to intercept a new one by applying an HMM-based method. Starting from a static set of multimodal ambiguity classes [10], DAMA can dynamically identify classes of ambiguity and, when an ambiguity does not belong to any one of the existing classes, it allows creating a new one according to the evolution of the language, or refining the identification of an already existing class (see Figure 5).

The classification process starts from the acquisition of the multimodal sentences and, then, from the extraction of their meaningful features. These features are used to build the knowledge (i.e., the observation sequences of the method for classifying the multimodal ambiguities). The associations between the observation sequences and the ambiguity classes are used to train the HMM-based method. Once trained, the HMM-based method enables the extraction of the ambiguity class from the ambiguous multimodal sentences. When it is not possible to associate an ambiguous sentence with a specific class, then, the method defines a new ambiguity class and a dialogue with experts starts in order to validate the new class. During the dialogue, the role of the experts is to discern if the new ambiguity class and the connected meaningful features are correctly defined or if the ambiguous sentence can be assigned to an already existing class. In the latter case, the system does not correctly associate the ambiguity to the connected class; therefore, the knowledge about the ambiguity class needs to be updated through the features contained in the sentence, as Figure 5 describes. The following section describes the meaningful extracted features that characterize classes.

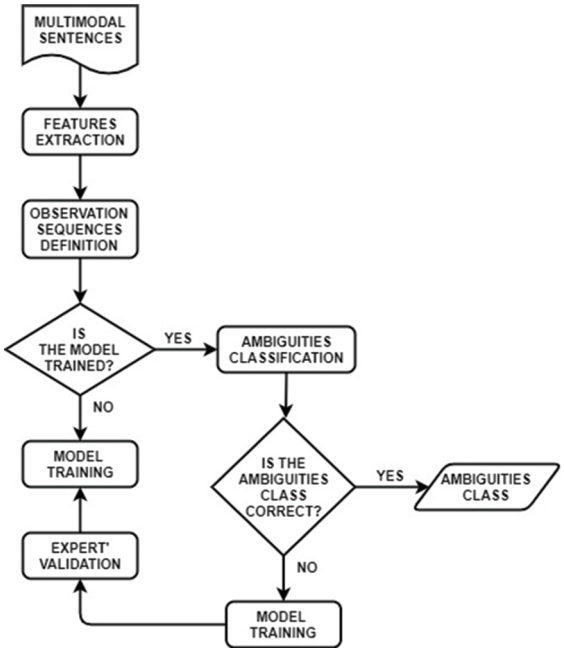

Figure 5 | Classifier flow chart.

6.1. The Observation Sequences

The information contained in the features vector f_t (defined in Section 4) of the multimodal sentence is used to create the observation sequence x_t as the ordered sequence of the pairs of features of the consecutive elements that compose the multimodal sentence:

$$x_t = \left[\left(E_r^v, E_r^{v+1}\right)\right] \quad (2)$$
$$with\ v \epsilon\ M\ and\ M = \{1..m\}$$
$$r \epsilon R\ and\ R = \{mod,\ repr,\ time,\ role,\ concept\}$$

We use pairs of features $\left(E_r^v, E_r^{v+1}\right)$ extracted from consecutive elements that compose the multimodal sentence, since these pairs of features can be associated to the values of variables (described in Section 5) that allow discriminating among the introduced ambiguity classes.

Considering the multimodal sentence of Figure 2, it consists of ten elements (i.e., E^1, E^2, E^3, E^4, E^5, E^6, E^7, E^8, $E^{9'}$, $E^{9''}$, $E^{10'}$, $E^{10''}$); the observation sequence x_t at time t is composed of pairs of the ten elements of the features vector, as Figure 6 shows.

As for the features vectors, the i^{th} observation symbol x^i_t of the observation sequence at time t may take all the values of the features of the elements included in the multimodal grammar [56]. All the values of these features compose the set O_k of values of the observation sequences x_t that are fed to the classifier.

As example, the observation sequence for the multimodal sentence of Figure 2 is:

x_t = [(speech, speech), ( "put"  "this"), (1, 2) (3, 5), (vb1, dt1, (put, this) ... (null, null), (null, null), (null, null), (nn1, nn2), (carrot, plate)]

The following section presents how to identify ambiguity classes by the ambiguity classifier.

	E^v_{mod} E^{v+1}_{mod}	E^v_{repr} E^{v+1}_{repr}	E^v_{time} E^{v+1}_{time}	E^v_{role} E^{v+1}_{role}	$E^v_{concept}$ $E^{v+1}_{concept}$
E^1 E^2	speech, speech	put this	(1,2), (3,5)	vb1, dt1	put, this
E^2 E^3	speech, gesture	this	(3,5), (2,5)	dt1, nn1	this, carrot
E^3 E^4	gesture, speech	on	(2,5), (5,6)	nn1, in1	carrot, on
E^4 E^5	speech, speech	on the	(5,6), (7,8)	in1, dt2	on, the
E^5 E^6	speech, speech	the plate	(7,8), (9,10)	dt2, nn2	the, plate
E^6 E^7	speech, speech	plate and	(9,10), (11,12)	nn2, cc1	plate, and
E^7 E^8	speech, speech	and eat	(11,12), (12,14)	cc1, vb2	and, eat
E^8 $E^{9'}$	speech, speech	eat it	(12,14), (14,15)	vb2, prp1	eat, carrot
$E^{9'}$ $E^{9''}$	speech, speech	it it	(14,15), (14,15)	prp1, prp1	carrot, plate
$E^{9''}$ $E^{10'}$	speech, null	it null	(14,15), null	prp1, nn1	plate, carrot
$E^{10'}$ $E^{10'}$	null, null	null, null	null, null	nn1, nn2	carrot, plate

Figure 6 | Pairs of elements that compose the observation sequence of the multimodal sentence of Figure 2.

6.2. The Classifier

The classifier addresses the problem to dynamically classify ambiguous multimodal sentences. Therefore, DAMA takes in input ambiguous multimodal sentences and gives as output the ambiguity class associated to the ambiguous multimodal sentence. In particular, the input domain is composed by the set of pairs of features (described in Section 4) that characterize the ambiguous multimodal sentences, while the output domain is composed by the set of ambiguity classes defined in Section 5.

Since we handles pairs of features of the consecutive elements (E_r^v, E_r^{v+1}) as observation sequences for our classifier, we need to use a model able to dynamically model sequences of complex data, and HMMs best fit this purpose. In addition, since these pairs of features of the consecutive elements (E_r^v, E_r^{v+1}) can be associated to the values of variables (described in Section 5), we model the set of hidden states as the set of values assumed by those variables:

$Q_k = \{N_1, CBR_1, EGS_1, PSG_1, SR_1, C_1, R_1, M_1, N_0, CBR_0, EGS_0, PSG_0, SR_0, C_0, R_0, M_0\}$.

Each hidden state produces a pairs of features (E_r^v, E_r^{v+1}); therefore, the sequence of hidden states and the sequence of pairs of features (observation sequences) have the same length (see Figure 7). Each ambiguity class can be characterized by sequences of these variables. For example, lexical ambiguity is characterized by two consecutive elements that have the same syntactic role (SR1), the same representation (R1) defined in the same modality (M1), but they refer to different concepts (Co). While, target ambiguity is characterized by two consecutive elements that have the same syntactic role (SR1), have two different representations (R0) defined in the same modality (M1), but refer to different concepts (C0). Therefore, the values of the transition probabilities among the hidden states Q_k characterize the different ambiguities classes. Therefore, for each ambiguity class, the hidden states need to be modelled as a single variable-value state, and the transition probabilities among these hidden states describe the specific peculiarities of the class.

In the proposed method each ambiguity class is described by its own model that corresponds to one specific HMM, as yet provided in other contexts such as for images classification [60,61]. The overall method includes the six different HMMs concerning the previously described classes, i.e., gap (ga), analytic (an), attachment (at), lexical (le), temporal-semantic (te), target (ta) and it also allows modelling new classes in the case that new ambiguity classes are intercepted, i.e., $ambiguity_{n1}, \ldots new_{nn}$, as Figure 8 shows.

For, each HMM_k (with k = gap, an, at, le, te, ta, $n_1, \ldots n_n$), the hidden states (Q_k) and the observation sequence O_k are represented by a dependency graph of the HMM that models the observation sequence O_k as shaded nodes and the hidden states Q_k as empty nodes (see Figure 7).

The sequence of these states variables allows discriminating the ambiguity class that appears in the multimodal sentence.

In order to clarify this topic, we again consider the example in Figure 2. Figure 7 shows the HMM dependency graph related only to the pairs of features of elements of the observation sequence ($E^{9'}$, $E^{9''}$) that generates the ambiguity (i.e., the pronoun defined by the speech element "  it" that may refer to both, the "carrot" or to the "plate" concept). Therefore, Figure 7 shows the connection between the hidden states (i.e., $M_1, R_1, CBR_1, SR_1, C_0$) and the pairs of features of the elements $E^{9'}$ and $E^{9''}$ (shaded nodes).

After describing the hidden states, the observation sequence and the connections among them, we continue to describe the general model giving the definitions of transition probability matrix, output probability matrix and initial distribution vector.

For, each HMM_k, the transition probability matrix A_k contains the probability to have transitions from the hidden state q_i to q_j that

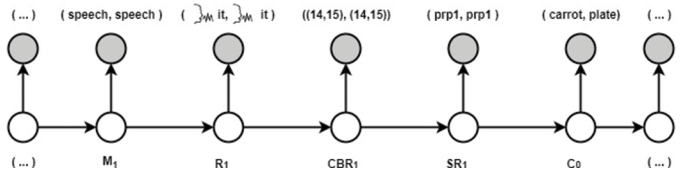

Figure 7 | Hidden Markov Model (HMM) related only to the pairs of features of the elements $E^{9'}$ and $E^{9''}$: empty nodes are hidden and shaded nodes are observed.

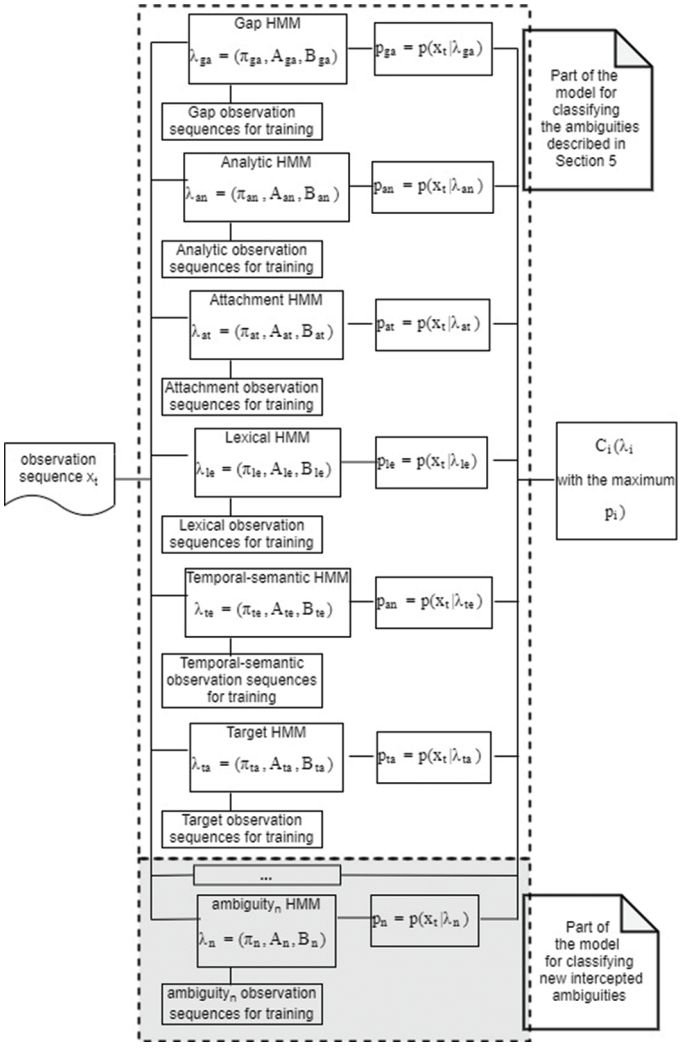

Figure 8 | The Hidden Markov Model (HMM)-based classifier for multimodal ambiguities.

belong to the set of hidden states Q_k of the model, as defined in the following formula.

$$A_k = \left[a_{ij}\right]_k \quad \text{with } a_{ij} = pr\left(q_{t+1} = s_j / q_t = s_i\right) \ \forall s_i, s_j \in Q_k \tag{3}$$

Moreover, the output probability matrix B_k defines the probability that each state $s_i \in Q_k$ will produce the observation sequence x_j from the set of observation sequences O_k.

$$B_k = \left[b_i\left(j\right)\right]_k \text{ with } b_i\left(j\right) = pr\left(x_t = o_j / q_t = s_i\right) \quad \forall s_i \in Q_k, \forall o_j \in O_k \tag{4}$$

Finally, π_k represents the initial distribution vector giving the probability that the state $s_i \in Q_k$ is the initial state of the sequence:

$$\pi_k = [\pi_i]_k \text{ with } \pi_i = pr\left(q_1 = s_i\right) \ \forall s_i \in Q_k \tag{5}$$

The parameters A_k, B_k, π_k allow specifying the HMM_k model λ_k of the ambiguity class k:

$$\lambda_k = (A_k, B_k, \pi_k) \tag{6}$$

The overall model is therefore composed by several HMMs, as Figure 8 shows, and the internal model of each HMM_k follows the structure of Figure 7. In particular, Figure 8 shows the HMM-based classifier. It is divided into two parts: the upper part represents the current state of knowledge in classifying ambiguities as described in Section 5; while the shaded part depicts the part of the model that allows adding new HMMs for a new intercepted ambiguity class. In this figure, different observation sequences O_k (defined as in Section 6.1), each one belonging to one of the possible classes, are used to train the specific HMM $\lambda_k = (A_k, B_k, \pi_k)$. When all the models are trained, a new observation sequence (x_t) can be classified. In order to associate the observation sequence (x_t) to one ambiguity class, the method computes the probability value $p(x_t/\lambda_k)$ for each model λ_k, and returns the most probable class (associated with λ_i) by finding the local maximum of the likelihood function for the probabilistic function:

$$C_i = \arg \max_{j=gap...ambuguity_{n_n}} p\left(x_t/\lambda_j\right) \tag{7}$$

The following section has the purpose of explaining more in detail the model training process, which has the purpose to learn the parameters vector of λ_k for each class k, and how the model works.

6.2.1. Training phase and operating principle

Since the purpose of the proposed model is to classify multimodal ambiguous multimodal sentence in one of the k ambiguity classes, we train k HMMS, one for each class, with the observation sequences (see the Section 6.1) corresponding to that class.

Each ambiguity class k has multiple observation sequences that are known to be generated by the same λ_k. Therefore, the training process of one λ_k, associated with the ambiguity class k, has the purpose to determine the λ_k parameters (i.e., transition probability matrix, output probability matrix, initial distribution vector) that fit the highest probability of generating the observed sequences, associated with the one connected ambiguity class.

The training process of each λ_k is defined by the following steps:

1. In the first step, the dataset is separated into n datasets (one data set for each multimodal ambiguity class);
2. In the second step, each λ_k is separately trained by the Baum-Welch algorithm [54,62] using the connected data set.

The training method is supervised since the ambiguities class label for each multimodal sentence is considered as known and used during the training phase.

When all the HMMs are trained, the overall model is able to identify the ambiguity class C_i to associate with the multimodal sentence. The class is associated by computing the likelihood probability $p(x_t/\lambda_i)$ by using the Forward–Backward algorithm [38,39]. The Forward–Backward algorithm computes the likelihood probabilities observation sequence x_t of the multimodal sentence respect to each λ_k. The comparison of the likelihood probabilities allows identifying the λ_i, associated with the highest likelihood probability. Considering a set of multimodal sentences and the class of ambiguities described in Section 5, the trained system allowed identifying a threshold likelihood probability value equal to 0.8 to associate any ambiguous multimodal sentence with the correct class of ambiguity. The threshold has been chosen performing the Receiver Operating Characteristic (ROC) curve analysis [63] on testing samples described in Section 7. Therefore, the likelihood probability must be higher than 0.8 (i.e., the threshold value) to be considered for assigning the multimodal sentence to the C_i.

When the probability value $p(x_t/\lambda_i)$ is lower than 0.8, it is not possible to associate an ambiguous sentence with a specific class, and then a new class needs to be evaluated. The following section describes how to perform the process for modeling a new class of multimodal ambiguity.

6.2.2. Case of new intercepted class

The need to have a dynamic method, able to classify new ambiguity, arises from the continuous evolution of the interaction language and from the purpose to extend the classification provided in [10] in order to treat the pragmatic ambiguity. For describing how the method manages new intercepted classes, we consider as starting state DAMA trained for classifying the six ambiguities (i.e., gap, analytic, attachment, lexical, temporal-semantic and target) described in Section 5. In particular, our purpose is to explain how DAMA works with the pragmatic ambiguity as the new ambiguity class to be defined.

The example of pragmatic ambiguity described in Section 6 (i.e., "put this on the plate and eat it" provided by speech and "the carrot" indicated by gesture) is given as input to DAMA trained for classifying gap, analytic, attachment, lexical, temporal-semantic and target ambiguity. DAMA processes the observation sequence x_t connected to the multimodal sentence (see Figure 6). Figure 9 shows that DAMA returns as output the probabilities that associate the observation sequence to each ambiguity class.

The highest probability is assigned to the target ambiguity, and its value is $p(x_t/\lambda_{le}) = 0.75$, and it is lower than the threshold value, which is 0.8. The obtained probability values reflect the similarity between the new one and the existing ambiguity classes. The ambiguity derives from a pronoun that is not directly referred to a

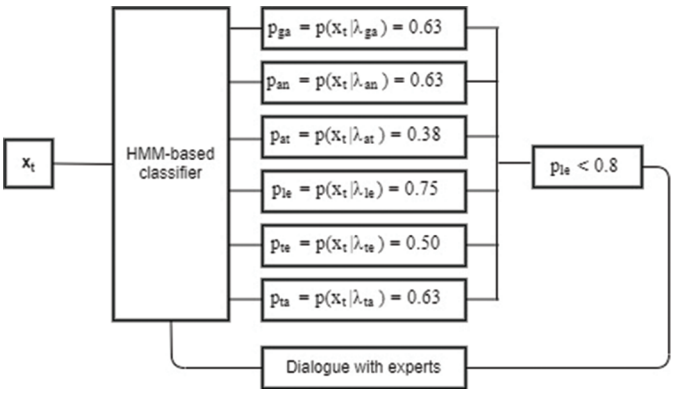

Figure 9 | Example of the case of new class intercepted.

concept and, as a consequence, the concept to which it refers is omitted. In addition, the pronoun may be connected to two different concepts cited in the sentence (carrot and plate). Since the highest returned probability is lower than the threshold value, it is not possible to associate the observation to one of the existing classes. In this case, the system enables a dialogue with the experts aiming to obtain their feedback to select one the following cases and refining the overall model:

- case 1: the ambiguity class has been not identified but it is one of the existing ones,
- case 2: a new ambiguity class needs to be identified.

In the first case, the observation sequence, derived from the multimodal sentence, is used to train and refine the HMM of the connected ambiguity class.

In the second case, a new class needs to be introduced and, the experts have to:

i. define a set of observation sequences extracted from multimodal sentences connected to the considered new class of ambiguity.

ii. identify which are the variables (see Section 5) that allow discriminating the new ambiguity class.

The defined set of observation sequences and the identified variables constitute the training set for the new class, which is added to the HMM-based classifier. In particular, for the sentence of Figure 2, the experts decide to define a new class because the model returned class with greater probability (0.75), but lower than the threshold (0.8).

In that example, we have two different interpretations (i.e., "put this carrot on the plate and eat the carrot" and "put this carrot on the plate and eat the plate") of the multimodal sentence that are caused by the information about the object referred to the pronoun "it" of the action "eat" that is missing; and the pronoun "it" may be inferred to two different elements (i.e., "carrot" and "plate").

The new class specification consists of the following steps:

1. To define a set the observation sequence.
2. To identify the discriminating variables (see Section 5) of the new one respect to the other existing classes.

- In the example of Figure 2, the multimodal sentence has two possible interpretations as the syntax-graph of the multimodal sentence in Figure 4 shows. Considering the Natural Language Processing (NLP) roles, the syntactic role *np* implies the *prp* syntactic role, connected to the pronoun "*it,*" and the *np* role connected to an element that have some features (E^{10}_{mod}, E^{10}_{repr}, E^{10}_{time}) that assumes the value null. This element can be associated to both the elements referred to "*carrot*" and the element referred to "*plate*." Therefore, there is an element that is omitted (N_1). Differently from the gap ambiguity the element, referred by the pronoun, may be connected (ESG_1) two different concepts (C_0) related to two different elements of the multimodal sentence (i.e., "*carrot*" defined by gesture and "*plate*" defined by speech).
- Therefore, there is one element of the multimodal sentence (SR_1, R_1, M_1) that has some features omitted (N_1) but may be referred (ESG_1) to two different concepts (C_0).

This approach implements a very flexible model since it allows adding extra HMMs without affecting the already trained models. In fact, each model for each ambiguities class is independently (from the models connected to the other classes) trained on its own training set, and it has no knowledge of any other models of other ambiguities classes and their training sets.

6.3. DAMA Implementation

The goal of this research was to develop a dynamic classification of multimodal ambiguities. In order to achieve this purpose, we applied an HMM-based approach implemented using *JAHMM* (code.google.com/archive/p/jahmm/downloads), that is a Java implementation of HMMs. In order to model the information needed for the classification, the following java libraries have been used:

- *JGraphX (*github.com/jgraph/jgraphx): this library is used in order to visualize the syntax-graph of the multimodal sentence;
- *Jdsl* (cs.brown.edu/cgc/jdsl/)*:* this library is used to create and manage complex data such as list, queue, tree, graph and priority queues; it is used in order to manage all structures that implies managing graphs;
- *Stanford CoreNLP* (stanfordnlp.github.io/CoreNLP/)*:* this is the java implementation of the Stanford Parser; it is applied for obtaining the syntactic tree connected with the sentence in natural language that represents the candidate interpretation of the multimodal sentence.

The following section provides the evaluation performance of the implemented model.

7. EVALUATION OF DAMA PERFORMANCE

The evaluation process consisted of two phases: the training phase and the test phase. The training phase has been performed as described in the section "*Training phase and operating principle*,"

and, subsequently, the trained model has been evaluated. The experiments, related with the evaluation process, were performed with training data and test data, and the size is chosen in order to have representative data [64]. The used data are extracted from 480 multimodal sentences containing the samples of ambiguities classes described [10] (i.e., gap, analytic, attachment, lexical, temporal-semantic and target) as well as samples of pragmatic ambiguities. These sentences were labeled to match them into the appropriate classes of ambiguities. We decide to equally split the entire data set for training and testing the model in order to have a significant number of samples for each class of ambiguities. In particular, the entire data were divided into two data sets, one for the training phase and the other for the test phase and both involved 240 multimodal sentences selected, as introduced in the section "*Training phase and operating principle.*" The test phase involved ambiguous multimodal sentences that were not used for training. This choice has been supported by the accuracy reached in the testing phase because the accuracy from 168 and 240 ambiguous multimodal sentences few grown, as Figure 13 shows. Therefore, improving the training set does not produce significant advantages. As a first step, we perform the test on DAMA trained for the six ambiguity classes (gap, analytic, attachment, lexical, temporal-semantic and target).

The Table 1 shows the generated confusion matrix of the ambiguity classification models based on multimodal sentences. The rows represent the number of true classifications made by model as gap, analytic, attachment, lexical, temporal-semantic and target. The columns represent the predicted classifications in the test data.

The testing phase consisted of providing three performance evaluation measures for each one of the trained ambiguity models:

- precision (P_i): that measures the fraction of the relevant instance (multimodal sentences that are correctly classified in the considered ambiguity class) among the retrieved instances (multimodal sentences that are classified in the considered class);
- recall (R_i): that measures the fraction of the relevant instance (multimodal sentences that are correctly classified in the considered ambiguity class) among all the total amount of the relevant instances (multimodal sentences that are associated to the considered class);
- specifity (S_i): that measures the proportion of no true classes that are correctly identified as such.

Precision, recall and specifity are all measures of relevance for the classification model. High precision means that the model returns most relevant instances than irrelevant ones, while high recall means that the model returns most of the relevant instances. Specifity quantifies the avoiding of no true classes that are classified as true, therefore, high specifity means a low type I error rate.

For each HMM_i trained for classify the ambiguity class *i with i*∈ *I and I = {Gap, Analytic, Attachment, Lexical, Temporal-semantic, Target}*

Those measures are defined as follows [65]:

$$P_i = \frac{\sum_{j=Gap}^{Target} x_{jj}}{\sum_{j=Gap}^{Target} x_{jj} + \sum_{j\neq Gap}^{Target} x_{ji}}$$

$$with\ j \epsilon J\ and\ J = \{Gap,\ Analytic,\ Attachment,\ Lexical,\ Temporal\text{-}semantic,\ Target\} \quad (8)$$

$$R_i = \frac{\sum_{j=Gap}^{Target} x_{jj}}{\sum_{j=Gap}^{Target} x_{jj} + \sum_{\substack{j=Gap \\ j\neq i}}^{Target} x_{ij}}$$

$$with\ j \epsilon J\ and\ J = \{Gap,\ Analytic,\ Attachment,\ Lexical,\ Temporal\text{-}semantic,\ Target\} \quad (9)$$

$$S_i = \frac{\sum_{i=Gap}^{Target}\sum_{\substack{j=Gap \\ j\neq i}}^{Target}\sum_{\substack{k=Gap \\ k\neq i}}^{Target} x_{jk}}{\sum_{i=Gap}^{Target}\sum_{\substack{j=Gap \\ j\neq i}}^{Target}\sum_{\substack{k=Gap \\ k\neq i}}^{Target} x_{jk} + \sum_{\substack{j=Gap \\ j\neq i}}^{Target} x_{ij}}$$

$$with\ j \epsilon J\ and\ J = \{Gap,\ Analytic,\ Attachment,\ Lexical,\ Temporal\text{-}semantic,\ Target\} \quad (10)$$
$$with\ k \epsilon K\ and\ K = \{Gap,\ Analytic,\ Attachment,\ Lexical,\ Temporal\text{-}semantic,\ Target\}$$

Table 1 presents the summary of the experiments and, in particular, provides the normalized multi-class confusion matrix of the ambiguity classification model performed on the 240 multimodal sentences associated to the 6 ambiguity classes (40 multimodal sentences for each ambiguity class). Table 1 displays the results of the evaluation parameters for all the ambiguity classification models and Figure 10 displays the rate of specifity, recall and precision for comparative analysis of all the different models. Main classification confusion was between analytic and attachment, and temporal-semantic and lexical and target due to the similarity between the features that characterize the ambiguity classes.

Table 1 | Confusion matrix of the ambiguity classification model.

		Predicted					
		ga	**an**	**at**	**le**	**te**	**ta**
True	**ga**	0,95	0,03	0,03	0,00	0,00	0,00
	an	0,03	0,88	0,05	0,00	0,03	0,00
	at	0,03	0,10	0,85	0,00	0,00	0,03
	le	0,00	0,00	0,03	0,93	0,03	0,03
	te	0,00	0,00	0,00	0,05	0,90	0,05
	ta	0,00	0,00	0,05	0,03	0,05	0,90

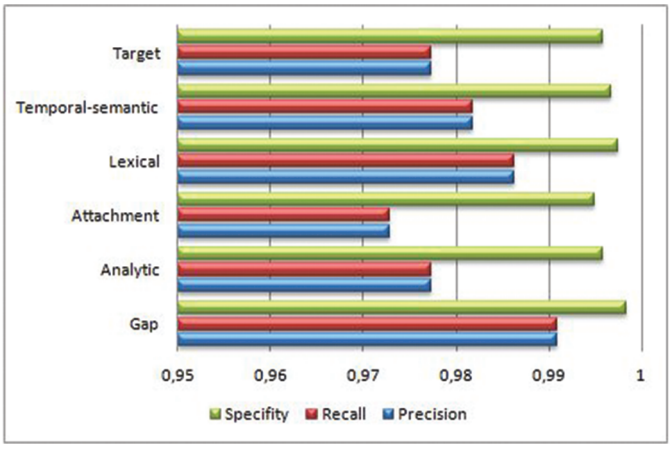

Figure 10 | Ambiguity recognition rates.

The classification performance of the overall method has been evaluated in terms of accuracy rate. We evaluate the overall accuracy of all the six models because it directly reflects the certainty of classification of outputs. In particular, accuracy is quantitatively measured as:

$$OverallAccuracy = \frac{\sum_{j=Gap}^{Target} x_{jj}}{N}$$

$$with\ j \epsilon J\ and\ J = \{Gap,\ Analytic,\ Attachment,\ Lexical,\ Temporal\text{-}semantic,\ Target\ \ and\ N\ is\ the\ total\ number\ of\ samples \tag{11}$$

The bigger is the accuracy the better is the result.

Figure 11 shows that the learning rate for six ambiguity classes improves when improving the amount of data in the training set. It shows that the accuracy rate grows from 62.5%, for the first 24 to 89.6%, for all the 240 ambiguous multimodal sentences of the training set.

This improvement is due to the fact that DAMA learns the connections among meaningful features of the multimodal sentences and the ambiguity classes by increasing the number of examples.

At a later time, the multimodal ambiguities, which are incorrectly classified by DAMA, have been validated by the experts and a new HMM (in that case the model for the co-reference ambiguity class) has been added to DAMA (Figure 12).

Figure 13 shows the comparison of the accuracy rate values obtained respectively for 6 (Figure 11) and 7 (Figure 12) classes of multimodal ambiguities.

Figure 13 shows that the accuracy of the classification process is improved by a greater number of multimodal ambiguous sentences in the training set and by increasing the number of the HMMs connected to the ambiguity classes from 6 to 7, because in this example a new class has been needed.

We tested the proposed approach by focusing on a single new ambiguity type as the used data were extracted from multimodal

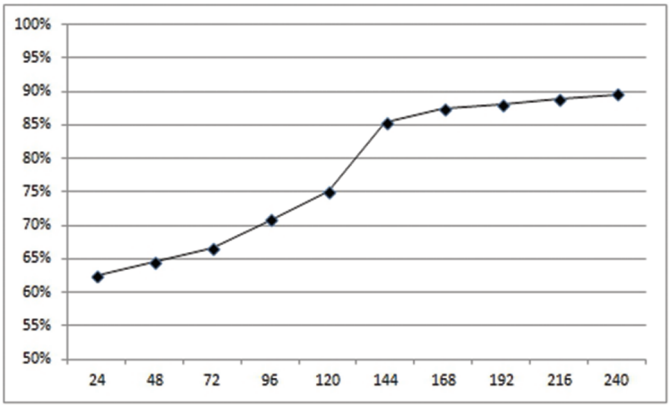

Figure 11 | Classification the overall accuracy rate for different amounts of data modelled in 6 classes of multimodal ambiguities.

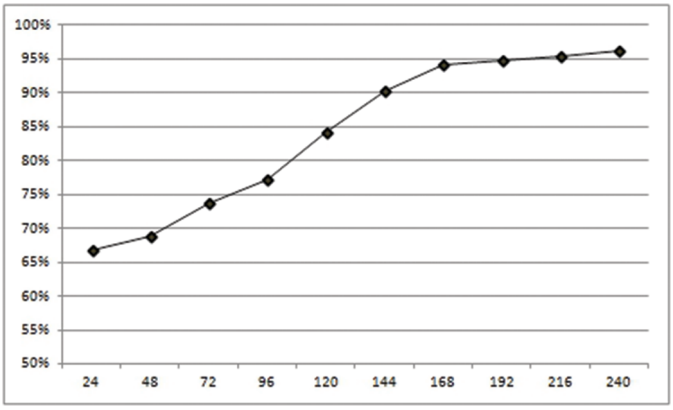

Figure 12 | Classification accuracy rate for different amounts of data modelled in 7 classes of multimodal ambiguities.

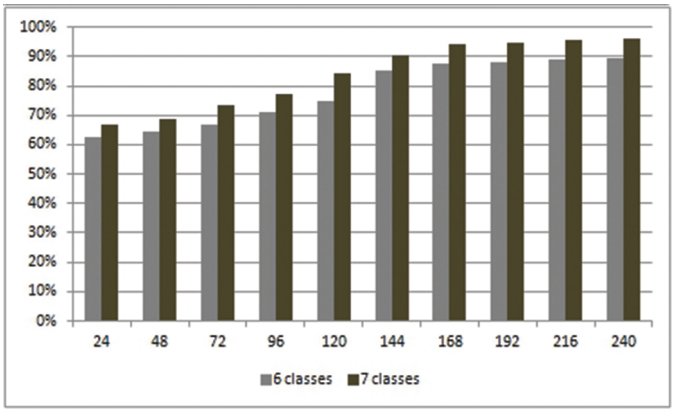

Figure 13 | Improvement of classification accuracy rate from 6 to 7 classes of multimodal ambiguities for different amounts of data.

sentences containing the samples of gap, analytic, attachment, lexical, temporal-semantic and target as well as samples of pragmatic ambiguities. The generalization capabilities of the method need an extension of the used dataset. This implies a wider multimodal corpus that we will build and validate. For these reasons, we will prove the generalization capabilities of our approach in a future work.

8. CONCLUSIONS

Since ambiguity issues are relevant in several disciplines (such as for writing, linguistics, philosophy, law, security and surveillance and human–machine interaction), the identification of the specific type of ambiguity has been addressed in this paper. In addition, the evolution of the interaction language and, so, the evolution of the language ambiguities is the need that had driven this work in order to define a classification method, which is able to learn and fit. This HMM-based method models ambiguous multimodal sentences according to a linguistic approach and starting from the classification method proposed in [10]. Moreover, DAMA has the ability to capture characteristics that are needed to distinguish ambiguity classes. As well as to distinguish between the classes, DAMA allows assessing whether if the ambiguity class is correctly identified or if a new ambiguity class needs to be defined. The proposed method allows introducing new classes by defining a specific set of observation sequences that are used to train the new class and refining the existing ones.

The values of the classification accuracy increased from 94.6% for the semantic ambiguities and 92.1% for the syntactic ambiguities for the static classification in [10], to 96.3% for the proposed method. This improvement is due to the use of a probabilistic model combined with a threshold. The flexibility is an important advantage for DAMA because extra classes can be added when the users need without affecting the trained HMMs. However, if too many classes are added, then the complexity is increased.

In future works, the classification process will be improved by testing if more information in the feature vectors during the training process allows improving the classification accuracy. This information will consider knowledge about the user model, the users' behavior, and the interaction history. This information will be investigated in order to analyze how it can be useful to refine the ambiguities classification for the different users. In addition, we will prove the generalization capabilities of our method extending the used dataset.

CONFLICT OF INTEREST

The authors declare they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

All authors contributed to the work. All authors read and approved the final manuscript.

Funding Statement

This research received no external funding.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the article.

REFERENCES

[1] S.T. Piantadosi, H. Tily, E. Gibson, The communicative function of ambiguity in language, Cognition. 122 (2012), 280–291.
[2] M. McLuhan, Q. Fiore, The Medium is the Massage, Random House, New York, 1967.
[3] M. Stacey, C. Eckert, Against ambiguity, Comput. Supp. Coop. Work. 12 (2003), 153–183.
[4] P.M. Aoki, A. Woodruff, Making space for stories: ambiguity in the design of personal communication systems, in Proceeding of CHI, Portland, OR, 2005, pp. 181–190.
[5] D.M. Berry, E. Kamsties, D.G. Kay, M.M. Krieger, From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity, Technical Report, University of Waterloo, Waterloo, ON, Canada, 2001.
[6] R.P. Futrelle, Ambiguity in visual language theory and its role in diagram parsing, in IEEE Symposium on Visual Languages, IEEE Computer Society, Tokyo, Japan, 1999, pp. 172–175.
[7] F. Favetta, M.A. Aufaure-Portier, About ambiguities in visual GIS query languages: a taxonomy and solutions, in Proceedings of the 4th International Conference on Advances in Visual Information Systems, Springer-Verlag, Lyon, France, 2000, pp. 154–165.
[8] M.C. Caschera, F. Ferri, P. Grifoni, The Management of Ambiguities, Visual Languages for Interactive Computing: Definitions and Formalizations, IGI Global, Hershey, PA, 2008, pp. 129–140.
[9] S. Oviatt, Multimodal interfaces, in: J.A. Jacko, A. Sears (Eds.), The Human-Computer Interaction Handbook, L. Erlbaum Associates Inc., Hillsdale, NJ, 2002, pp. 286–304.
[10] M.C. Caschera, F. Ferri, P. Grifoni, From modal to multimodal ambiguities: a classification approach, JNIT. 4 (2013), 87–109.
[11] M.C. Caschera, F. Ferri, P. Grifoni, An Approach for Managing Ambiguities in Multimodal Interaction, OTM 2007 Ws, Part I, LNCS 4805, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 387–397.
[12] M.C. Caschera, Interpretation methods and ambiguity management in multimodal systems, in: P. Grifoni (Eds.), Multimodal Human Computer Interaction and Pervasive Services, IGI Global, Hershey, PA, 2009, pp. 87–102.
[13] M.C. Caschera, A. D'Ulizia, F. Ferri, P. Grifoni, Towards evolutionary multimodal interaction, in OTM 2012 Workshops Proceedings, Lecture Notes in Computer Science, Springer-Verlag, Rome, 2012, vol. 7567, pp. 608–616.
[14] S.D. Qiu, English Ambiguity, The Commercial Press, Beijing, China, 1998.
[15] L. Tang, Functions of pragmatic ambiguity on the English joke, Int. J. Engl. Lang. Lit. Humanit. 4 (2016), 49–58. http://ijellh.com/OJS/index.php/OJS/article/view/1424/1381
[16] G. Vigliocco, P. Perniss, D. Vinson, Language as a multimodal phenomenon: implications for language learning, processing and evolution, Philos. Trans. R Soc. B. 369 (2014), 1–7. http://rstb.royalsocietypublishing.org/content/royptb/369/1651/20130292.full.pdf
[17] E. Jablonka, S. Ginsburg, D. Dor, The co-evolution of language and emotions, Philos. Trans. R. Soc. B Biol. Sci. 367 (2012), 2152–2159.
[18] B. Gleich, O. Creighton, L. Kof, Ambiguity Detection: Towards a Tool Explaining Ambiguity Sources, vol. 6182, Springer-Verlag, Berlin, Heidelberg, 2010.

[19] V. Hegde, Multi-Perspective Comparative Study: Common Context Based knowledge integration in Word Sense Disambiguation for Information Retrieval, PhD thesis in Computer Science and Engineering from Avinashilingam University, Coimbatore, India, 2012. https://shodhganga.inflibnet.ac.in/bitstream/10603/72398/8/vhegdevinay_intro.pdf

[20] H.C. Maurya, P. Gupta, N. Choudhary, Natural language ambiguity and its effect on machine learning, Int. J. Mod. Eng. Res. 5 (2015), 25–30. http://www.ijmer.com/papers/Vol5_Issue4/Version-1/D0504_01-2530.pdf

[21] M.C. Caschera, F. Ferri, P. Grifoni, InteSe: an integrated model for resolving ambiguities in multimodal sentences, IEEE Trans. Syst. Man Cybern. Syst. 43 (2013), 911–931.

[22] M.C. Caschera, F. Ferri, P. Grifoni, Ambiguity detection in multimodal systems, in: S. Levialdi (Ed.), Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2008), Napoli, Italy, 2008, pp. 331–334.

[23] M.C. Caschera, F. Ferri, P. Grifoni, Multimodal interaction systems: information and time features, Int. J. Web. Grid Serv. 3 (2007), 82–99.

[24] H. Yang, A.n.d. Roeck, V. Gervasi, A. Willis, B. Nuseibeh, Analysing anaphoric ambiguity in natural language requirements, Requir. Eng. 16 (2011), 163–189.

[25] D.M. Berry, E. Kamsties, M.M. Krieger, From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity, Technical Report, University of Waterloo, Waterloo, ON, Canada, 2003. https://cs.uwaterloo.ca/~dberry/handbook/ambiguityHandbook.pdf.

[26] N. Kiyavitskaya, N. Zeni, L. Mich, D.M. Berry, Requirements for tools for ambiguity identification and measurement in natural language requirements specifications, Requirements Eng. 13 (2008), 207–239.

[27] D.M. Berry, R. Gacitua, P. Sawyer, S.F. Tjong, The case for dumb requirements engineering tools, in: B. Regnell, D. Damian (Eds.), Requirements Engineering: Foundation for Software Quality. REFSQ 2012. Lecture Notes in Computer Science, vol. 7195, Springer, Berlin, Heidelberg, 2012, pp. 211–217.

[28] A.K. Massey, R.L. Rutledge, A.I. Anton, P.P. Swire, Identifying and classifying ambiguity for regulatory requirements, in 2014 IEEE 22nd International Requirements Engineering Conference (RE), 2014, pp. 83–92.

[29] S. Gong, C.C. Loy, T. Xiang, Security and surveillance, in: T. Moeslund, A. Hilton, V. Krüger, L. Sigal (Eds.), Visual Analysis of Humans, Springer, London, 2011, pp. 455–472.

[30] T.S. Lim, W.Y. Loh, Y.S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Mach. Learn. 40 (2000), 203–228.

[31] P.N. Tan, M. Steinbach, V. Kumar, Classification: basic concepts, decision trees, and model evaluation, in: Introduction to Data Mining, Addison-Wesley Longman Publishing Co., Boston, MA, 2006, pp. 145–205. https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf

[32] A. Skowron, H. Wang, A. Wojna, J. Bazan, Multimodal classification: case studies, in: J.F. Peters, A. Skowron (Eds.), Transactions on Rough Sets V. Lecture Notes in Computer Science, vol. 4100 Springer, Berlin, Heidelberg, 2006, pp. 224–239.

[33] R. Quinlan, Improved use of continuous attributes in C4.5, J. Artif. Intell. Res. 4 (1996), 77–90.

[34] A.G. Wojna, Analogy-based reasoning in classifier construction, in: J.F. Peters, A. Skowron (Eds.), Transactions on Rough Sets IV, Lecture Notes in Computer Science, vol. 3700, Springer, Berlin, Heidelberg, 2005, pp. 277–374.

[35] J.G. Bazan, M. Szczuka, J. Wróblewski, A new version of rough set exploration system, in Proceeding of RSCTC'2002, Lecture Notes in Artificial Intelligence 2475, Springer-Verlag, Heidelberg, 2002, pp. 397–404.

[36] L. Ben-Gal, Bayesian networks, in: F. Ruggeri, F. Faltin, R. Kenett (Eds.), Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons, 2007.

[37] C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc. 2 (1998), 1–47.

[38] G.P. Zhang, Neural networks for classification: a survey, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 30 (2000), 451–462.

[39] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE. 77 (1989), 257–285.

[40] T.M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, New York, NY, 1997, pp. 432. https://dl.acm.org/doi/book/10.5555/541177

[41] A. An, Learning classification rules from data, Comput. Math. Appl. 45 (2003), 737–748.

[42] Z. Zenn Bien, L. Hyong-Euk, D. Jun-Hyeong, K. Yong-Hwi, P. Kwang-Hyun, Y. Seung-Eun, Intelligent interaction for human-friendly service robot in smart house environment, Int. J. Comput. Int. Sys. 1 (2008), 77–93.

[43] J. Cheng, R. Greiner, J. Kelly, D. Bell, W. Liu, Learning Bayesian networks from data: an information-theory based approach, Artif. Intell. 137 (2002), 43–90.

[44] L. Kessous, G. Castellano, G. Caridakis, Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis, J. Multimodal User In. 3 (2010), 33–48.

[45] S.B. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica. 31 (2007), 249–268. http://www.informatica.si/index.php/informatica/article/view/148/140

[46] M. Malcangi, P. Grew, Evolving connectionist method for adaptive audio visual speech recognition, Evol. Syst. 8 (2017), 85–94.

[47] M. Antal, Speaker independent phoneme classification in continuous speech. Studia Univ. Babes-Bolyal Inform. 49 (2004), 55–64. http://www.cs.ubbcluj.ro/~studia-i/contents/2004-2/6-Antal.pdf

[48] N. Liu, B. C. Lovell, Gesture classification using hidden Markov models and Viterbi path counting, in: C. Sun, H. Talbot, S. Ourselin, T. Adriaansen (Eds.), Proceedings of the Seventh Biennial Australian Pattern Recognition Society Conference, The Seventh Biennial Australian Pattern Recognition Society Conference, Sydney, Australia, 2003, pp. 273–282. https://espace.library.uq.edu.au/view/UQ:10700

[49] M.Y. Chen, A. Kundu, J. Zhou, Off-line handwritten word recognition using a hidden Markov model type stochastic network, IEEE Trans. Pattern Anal. Mach. Intell. 16 (1994), 481–496.

[50] S. Argyropoulos, K. Moustakas, A. Karpov, O. Aran, D. Tzovaras, T. Tsakiris, G. Varni, B. Kwon, Multimodal user interface for the communication of the disabled, J. Multimodal User In. 2 (2008), 105–116.

[51] C. Lu, M.S. Drew, J. Au, Classification of summarized videos using hidden markov models on compressed chromaticity signatures, in ACM Multimedia, Ottawa, Canada, 2001, pp. 479–482.

[52] N. Novielli, HMM modeling of user engagement in advice-giving dialogues, J. Multimodal User In. 3 (2010), 131–140.
[53] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, M. Meteer, Dialogue act modeling for automatic tagging and recognition of conversational speech, Comput Linguist. 26 (2000), 3.
[54] D.P. Twitchell, M. Adkins, J.F. Nunamaker, J.K. Burgoon, Using speech act theory to model conversations for automated classification and retrieval, in Proceeding of the 9th International Working Conference on the Language-Action Perspective on Communication Modeling, New Brunswick, NJ, 2004, pp. 121–130.
[55] A. Martalo, N. Novielli, F. de Rosis, Attitude display in dialogue patterns, in Proceedings of AISB'08, Symposium on 'Affective Language in Human and Machine', The Society for the Study of Artificial Intelligence and Simulation of Behaviour, Aberdeen, UK, 2008. https://www.academia.edu/7595837/Attitude_Display_in_Dialogue_Patterns
[56] H.S. Chen, W.J. Tsai, Incorporating frequent pattern analysis into multimodal HMM event classification for baseball videos, Multimed. Tools Appl. 75 (2016), 4913–4932.
[57] A. D'Ulizia, F. Ferri, P. Grifoni, Generating multimodal grammars for multimodal dialogue processing, IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 40 (2010), 1130–1145.
[58] M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, Building a large annotated corpus of English: the Penn Treebank, Comput. Linguist. 19 (1994), 313–330.
[59] T. Tung, R. Gomez, T. Kawahara, T. Matsuyama, Multiparty interaction understanding using smart multimodal digital signage, IEEE Trans. Hum. Mach. Syst. 44 (2014), 625–637.
[60] M. Mouret, C. Solnon, C. Wolf, Classification of images based on hidden Markov models, in IEEE Workshop Content Based Multimedia Indexing, Chania, Greece, 2009, pp. 169–174.
[61] H. Josinski, D. Kostrzewa, A. Michalczuk, A. Switonski, K.W. Wojciechowski, Feature extraction and HMM-based classification of gait video sequences for the purpose of human identification, Vision Based Syst. Appl. 481 (2013), 233–245.
[62] T. Benesch, The Baum-Welch algorithm for parameter estimation of Gaussian autoregressive mixture models, J. Math. Sci. 105 (2001), 2515–2518.
[63] A. Tharwat, Classification assessment methods, Appl. Comput. Inform. (2018), ISSN2210-8327.
[64] P. Glauner, P. Valtchev, R. State, Impact of biases in big data, in Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018), Bruges, Belgium, 2018. arXiv:1803.00897
[65] C. Manliguez, Generalized Confusion Matrix for Multiple Classes, 2016.