Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 480–489

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

How are you? How a Robot can Learn to Express its own Roboceptions

Irene Trifirò[a], Agnese Augello[b], Umberto Maniscalco[b], Giovanni Pilato[b,*], Filippo Vella[b], Rosa Meo[a]

[a]*Dipartimento di Informatica, Università degli Studi di Torino , Corso Svizzera 185, I-10149 Torino (ITALY)*
[b]*Institute for high performance computing and networking, National Research Council (CNR), Via ugo la Malfa 153, 90146 Palermo, Italy*

Abstract

This work is framed on investigating how a robot can learn associations between linguistic elements, such as words or sentences, and its bodily perceptions, that we named "roboceptions". We discuss the possibility of defining such a process of an association through the interaction with human beings. By interacting with a user, the robot can learn to ascribe a meaning to its roboceptions to express them in natural language. Such a process could then be used by the robot in a verbal interaction to detect some words recalling the previously experimented roboceptions. In this paper, we discuss a Dual-NMT approach to realize such an association. However, it requires adequate training corpus. For this reason, we consider two different phases towards the realization of the system, and we show the results of the first phase, comparing two approaches: one based on the Latent Semantic Analysis paradigm and one based on the Random Indexing methodology.

Keywords: Human-Robot Interaction; Roboceptions; Language Grounding; Latent Semantic Analysis (LSA); Dual-NMT

1. Introduction

Language learning is a complex but at the same time fascinating research topic. Experiments support the theory that language acquisition is strongly interconnected with probabilistic learning. There are many models attempting to represent this process. However, the structure of these mechanisms is controversial. Computational models can serve as a test bench for probabilistic learning and appraise model efficacy. Recently, there has been an increasing interest in considering an "embodied use of language" in human-robot interaction [23]. As stated in [6], "the dialogue and communication between robots and people should be designed not on a simple manipulation of detached symbols alone but on the relationship with the physical interaction between the robot and the surrounding world in which

* Corresponding author
E-mail address: giovanni.pilato@icar.cnr.it

1877-0509
10.1016/j.procs.2020.08.050

the language is grounded". In this context, many works rely on a developmental approach, defining the artificial language acquisition models taking inspiration from the developmental mechanisms observed in children [5] [20]. Language learning in children begins in a phase, around 10-12 months, when they have made enough perceptual and motor experiences that led them to collect a fair amount of pre-linguistic knowledge about physical and perceptual properties of the environment [12] [13] [4] [16]. This allows them to associate words to the observed and manipulated objects and to the most salient events [12].

The grounding of words is the first mechanism to associate words to their meaning since they are connected with perceivable entities. A second mechanism, altogether widely studied, is the reference to a linguistic context, that allows children to acquire new words by putting them in relations to other words [12].

A research field named SER (Symbol Emergence in Robotics) focuses on how a robot can learn words and their meaning by considering multimodal (visual auditory and in some cases also tactile) sensory information [27] [22] [17] [7]. There is also an increasing attention in modeling the behavior of robots according to their internal states. In [10][3], it has been described a bio-inspired somatosensory system for a humanoid robot proposed to simulate the processes under which human beings perceive pleasant and unpleasant physical sensations from the stimulation of their sensory receptors. Employing the somatosensory system, embedded in a cognitive architecture, a robot can roughly classify its sensations, that we named *roboceptions*. In [1], it is proposed a reasoning process letting a robot to conceptualize some sort of "synthetic" pain. The aim is to interpret, in some sense, human pain that can occur during collaborative tasks and as a consequence, that can generate an empathetic response. To implement this synthetic pain, the authors consider joint restriction regions, where each region is associated with a level of pain.

In this paper, we propose a mechanism of association of linguistic elements for a robot. We refer to the linguistic elements, such as words or sentences, with a robot internal perceptions. Starting from the arrangement of the stimuli in roboceptions, we discuss the definition of a conceptualization process mingled with the grounding of roboceptions. Words related to painful or well-being states are learned through a grounding process on the roboceptions provided by the somatosensory system. In our approach, the process by which the robot learns the language to describe its roboceptions is based on the interaction with a human-being that, similarly to a mother, explains to the robot how it can express its sensations. The mechanism is based on two phases: the first one is aimed at building a parallel corpus of sentences expressed both in an elementary language, that we named "Robolang", expressing the current status of the robot, and the corresponding Italian language. Once the parallel corpus is large enough, a second phase, based on Dual Neural Machine translation, can start. This second phase realizes the grounding language on roboception process. In this paper we illustrate the whole idea, that we have completely implemented, and we show the results of the first phase by comparing two approaches: one based on the Latent Semantic Analysis paradigm which exploits a set of pre-trained word vectors in Italian language and one based on the Random Indexing methodology, which does not require any *a priori* knowledge and builds the sub-symbolic representation of words from scratch.

2. Robotic somatosensory system

The body sensations of human beings are felt through a complex system of receptors, nerve fibers, and parts of the brain cortex that neuro-scientists call somatosensory system [21].

In [10], we have proposed an artificial somatosensory system that is inspired to the biological one, simulating how physical sensations are perceived by human beings, to fulfill a function similar to the human somatosensory system. However, in order to stress the obvious differences with the human system, we refer to these "sensations" with the term "roboceptions". The artificial somatosensory system is equipped with two paths: an ascending and a descending path. The biological receptors are replaced by the sensors of the robot. Nerve fibers are represented by the communication channels. The somatosensory cortex is made up of models that emulate the formation of sensations.

Every single roboception is obtained through the use of a special soft sensor [10]. The sensors of the robot constitute the input of the soft sensor, and their measurements are sampled at a suitable frequency for the type of roboception to be generated. The block represented by the gears implements the computational model that generates the roboception. All this concerns the ascending path. Instead, the descending path is represented by the inhibition and modulation blocks. The first one directly inhibits the stimulus, not even making it reach the computational model. The second one acts on the output and changes its value.

3. Grounding language on roboceptions through dual learning for Machine Translation

We consider that roboceptions can be expressed with a synthetic language, that we named *"Robolang"* representing the robot status. Roboceptions are coded, at a low level, by a vector representation. The robot, by using *"RoboLang"*, is able to manifest its basic needs, as occurs in a baby, when he is capable of expressing sensations with a very simple and limited set of words or onomatopoeic expressions. Similarly, *"RoboLang"* is composed of a set of elementary statements that have the form of a triplet: *(RoboceptionType, Location, Value)* where:

- *RoboceptionType* is the type of sensation; it can assume one of the following values: *CurrentPain, TemperaturePain, CaressFeeling, BumperPain, Anxiety, Exertion.*
- *Location* $\in \{0, 1, 2, \cdots, N\}$, each number identifying one of the N physical positions of the robot soft sensors. The value 0 indicates that the Roboception regards the whole robot and it has not a specific localization.
- *Value* $\in$ {*VeryLow, Low, Normal, High, VeryHigh*}, corresponding to the linguistic quantization of the robot soft sensors output value. If the sensor gives a binary output, the set of possible values are *VeryHigh* and *VeryLow*, associated to the values "True" and "False" respectively.

The statement is expressed only if a specific configuration of interest is triggered. Each specific configuration, at present, is hardwired in the somatosensory system (i.e., a set of thresholds have been set). In the future, these configurations will be learned by the robot in an autonomous manner.
The statements can also be concatenated, indicating the co-occurrence, during a given time interval, of multiple values of different relevant roboceptions at a given time. The *RoboLang* sentences represent the basic expression of the robot state or its primitive needs. As an example, these simple *RoboLang* "sentences" are possible:

- *(CurrentPain 2 High), (Energy 1 VeryLow)*
- *(Energy 1 High), (CaressFeeling 3 VeryHigh), (TemperaturePain 3 High)*

The idea illustrated in this paper is to link the simple expressions in *RoboLang* with more sophisticated sentences expressed in natural language.
The consequences of the approach are twofold: to let the robot learn how to better express itself naturally, and, vice-versa: to let the robot being capable of "understanding" and "internalizing" the sensations that a human being communicates through the use of natural language.
Since, basically, we have two languages to bridge (the robot and the human language), the grounding of roboceptions can be treated as a machine translation problem.

One of the most popular approaches to perform automatic translation between languages is the Neural Machine Translation (NMT). However, one of the big drawbacks of this kind of techniques is that they require a huge amount of bilingual sentences for training, while, in our case, the interaction between the human and the robot is very limited. This leads to a lack of training data.
To overcome this problem, a more promising approach can be used. In particular, in [11] it has been presented a dual-learning mechanism, named dual-NMT, which can be used on an NMT system to automatically learn from unlabeled data by exploiting a dual-learning game. The advantage of this system is that it needs only 10% of bilingual data for a warm start, reaching an accuracy comparable to traditional NMT.

Even applying the Dual-NMT approach, the number of sentences that are needed for training the neural network is relatively high. Hence a procedure that can help in a first phase is building a "parallel" RoboLang-Italian dataset. In this paper, we suggest a procedure which exploits the Latent Semantic Analysis (LSA) paradigm [15] and the word embeddings coding [9] for building such a parallel sentences dataset. Then, once such a dataset will be built, it could be used in a second phase for training a Dual-NMT approach.
Our goal is to lead the robot to eventually learn how to describe its own sensations with the result that the robot will seem more convincing and more "human". Furthermore, when the robot will listen someone expressing a feeling, it will be able to "computationally understand" the described sensation.
The whole process is sketched in Figure 1.

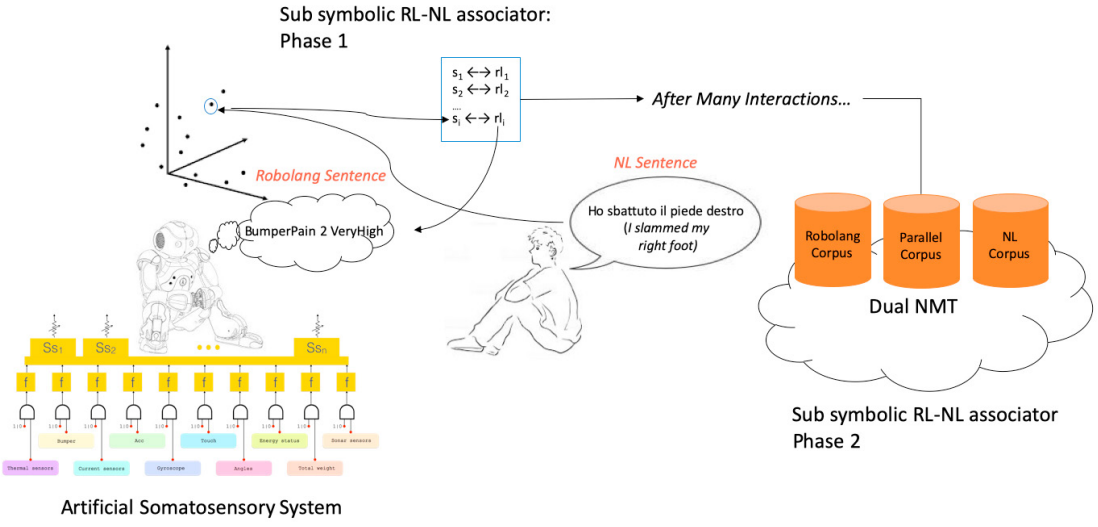

Fig. 1. Associations between RoboLang (RL) and Natural Language (NL) sentences in the two phases

3.1. Phase 1: Roboception grounding and improvement of the parallel corpus

Phase 1 of the grounding process exploits a sub-symbolic representation of sentences expressed in Natural Language and of sentences expressed in RoboLang.

During this phase, the interaction between human and robot is very limited, and the robot starts to learn expressing basic roboceptions. However, as long as the interaction is going on, new sentences either in Natural Language or in RoboLang are acquired and inserted in a parallel sentences dataset. In such a manner, the robot can autonomously learn new manners of expressing its roboceptions being also capable of recognizing and "properly" naming them as a human could do.

Let us suppose we have a RoboLang - Natural language corpus of n pairs of sentences $(r_1, s_1), (r_2, s_2), ..., (r_n, s_n)$ such that s_i is a natural language sentence and r_i is its RoboLang translation. Starting from these pairs we generate a lookup table where each entry is a pair (r_i, s_i).

Each one of these sentences is associated with a word vector which represents it. We used and compared two different approaches to generate the word vectors: an LSA-Word2Vec based approach and a pure Random Indexing approach. In both cases, the translation process works as follows. Let us consider a sentence s in Natural Language: we want to translate it in RoboLang. The translation process from RoboLang to Natural Language is completely analogous. First, we search for a pair (r_i, s_i) in the lookup table such that $s_i = s$. If such a pair exists, the translation of s is straightforward: r_i. Otherwise, we search for a pair (r_i, s_i) such that s_i is the most semantically similar sentence to s. The semantic similarity between two sentences is computed by cosine similarity between their word vectors. Let us suppose $\mathbf{v}$ and $\mathbf{v}_i$ are the word vectors of s and s_i, their similarity score is computed as $sim(s, s_i) = \frac{\mathbf{v}\mathbf{v}_i}{||\mathbf{v}||||\mathbf{v}_i||}$.

If the highest similarity score is less than a threshold experimentally fixed, the robot cannot translate the sentence and, therefore, it asks the user its meaning. The same happens when the robot experiments a totally new roboception. Once the bilingual data become sufficiently large, the grounding process will be tackled by a Dual-NMT approach.

3.1.1. LSA-Word2Vec based approach

In the LSA-Word2Vec approach, we used a Word2Vec model to embed the sentences expressed in Natural Language. Word2Vec is a predictive model for learning vector representations of words in very large corpora [19] [18]. In particular, we used the pre-trained vectors provided by FastText [9].

Since we do not have an already available and effective RoboLang word embedding, we decided to exploit the Latent Semantic Analysis (LSA) technique to obtain a sub-symbolic encoding of RoboLang sentences. LSA is a well-founded and well-known vector space method for extracting and representing latent concepts of a corpus of texts. It is

capable of emulating many cognitive phenomena [15][14]; furthermore, it is also a general approach that can be used on any dyadic domain [24] [8] and that has successfully been used for inducing data-driven bilingual dictionaries [28].

Let us suppose we have a corpus formed by M words and N documents. We build a $(M \times N)$ matrix whose rows are associated with words and whose columns are associated with documents; the *(i,j)-th* cell of $\mathbf{A}$ represents the frequency of the *i-th* word in the *j-th* document. Then we decompose this matrix, called word-document matrix, using the truncated singular value decomposition (T-SVD) in order to obtain an approximation of $\mathbf{A}$: $\hat{\mathbf{A}} \approx \mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U}$ is the $(M \times k)$ left singular matrix, $\mathbf{V}$ is the $(N \times k)$ right singular matrix and $\mathbf{\Sigma}$ is a $(k \times k)$ diagonal matrix with the top-k singular values of $\mathbf{A}$ in the decreasing order. The column vectors of $\mathbf{U}$ and the column vectors of $\mathbf{V}$ are orthogonal and form a k-dimensional basis: words, represented as $\mathbf{u}_i\mathbf{\Sigma}$, are embedded in the vector space formed by $\mathbf{V}$ column vectors; documents, represented as $\mathbf{v}_j\mathbf{\Sigma}$, are embedded in the vector space formed by $\mathbf{U}$ column vectors. In this manner, we consider the top-k most important latent semantics of our corpus.

In order to get a sub-symbolic representation of RoboLang words and sentences, let us consider the RoboLang sentences of our bilingual corpus, i.e. $r_1, r_2, ..., r_n$. Let us suppose that they are formed by m RoboLang distinct words; we denote them by $w_1, w_2, ..., w_m$. We construct a binary word-document matrix $\mathbf{A}$ whose elements indicate if a particular RoboLang sentence contains a particular RoboLang word, that is:

$$a_{ij} = \begin{cases} 1, & \text{if } w_i \text{ in } r_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Then we apply the Latent Semantic Analysis (LSA) technique to the matrix $\mathbf{A}$ in order to generate the k-rank approximated matrix $\hat{\mathbf{A}}$ (see eq. 3.1.1). In this manner the j-th RoboLang sentence is embedded as $\mathbf{d}_j = \mathbf{v}_j\mathbf{\Sigma}$.

When we have a new sentence r in RoboLang to translate, first we code it as a column of $\mathbf{A}$; we denote this column by $\mathbf{d}$. Then we embed $\mathbf{d}$ in the vector space of $\hat{\mathbf{A}}$ as: $\mathbf{d}_r = \mathbf{d}^T\mathbf{U}\mathbf{\Sigma}^{-1}$

If there is at least a new RoboWord in r, i.e., the robot experiments a new roboception, the grounding procedure recomputes a new word-document matrix $\mathbf{A}$ and applies again the LSA technique.

3.1.2. Pure Random Indexing approach

In the Pure Random Indexing approach, we exploited the Random Indexing (RI) technique to obtain a sub-symbolic representation of both RoboLang and Natural Language sentences starting from scratch.

RI is an incremental vector space model, presented in [26] as an alternative to LSA-like models. While LSA first constructs completely the co-occurrence matrix and then extracts context vectors, RI first accumulates context vectors and then collects these as rows of the co-occurrence matrix.

Let us suppose we have a corpus made up of n sentences, $s_1, s_2, ...s_n$, and of m words, $w_1, w_2, ..., w_m$. For each word w_i we randomly generated a unique d-dimensional vector that we call "index vector". Index vectors consist of thousands of elements set to 0, with a small number of randomly distributed elements set to +1s and -1s. To each word we associate a second vector, that we call context vector. Context vectors are computed by scanning our corpus with a fixed window. Each time a word w_i occurs, we added the index vectors of the words in the window to the context vector of w_i.

Let us consider the RoboLang sentences of our bilingual corpus, i.e. $r_1, r_2, ..., r_n$ and suppose that they are formed by m RoboLang distinct words; we denote them by $w_1, w_2, ..., w_m$. The encoding of sentences expressed in Natural Language is totally analogous. First we generate the index vectors of RoboLang words, i.e. $idx_1, idx_2, ..., id_m$. We associate to each word a context vector of 0, i.e. $cxt_1, cxt_2, ..., cxt_m$. Then we scan each sentence r_i with a k-dimensional window and update the context vectors of each word of r_i. In this manner we represent each sentence r_i with a vector v_i given by the mean of the context vectors of its words.

Let us suppose we have a new sentence r in RoboLang to translate. First we generate the index vector of each new word of r. Then we update the context vectors by scanning r with the k-dimensional window. We compute the word vector of r by averaging the context vectors of its words.

3.2. Phase 2: Dual-NMT

Once the parallel language corpus has become sufficiently large, a Dual-NMT approach can be exploited to realize the grounding language on roboceptions.

In the following we briefly recap the Dual-NMT procedure: let D_A and D_B be two monolingual corpora for languages A and B respectively. We have two weak translation models, from language A to language B and vice-versa; we denote them with $P(.|s;\Theta_{AB})$ and $P(.|s;\Theta_{BA})$ respectively. Also, we need two well-trained language models, $LM_A(.)$ and $LM_B(.)$, each of which returns a real value to indicate how a sentence is natural in its own language.

Let us suppose we start from a sentence s in D_A. Performing a beam search, we generate the top-K most probable middle translations of s according to the translation model $P(.|s;\Theta_{AB})$; let those be $s_{mid} = s_{mid,1}, s_{mid,2}, ..., s_{mid,K}$. For each middle translation $s_{mid,k}$, we compute its language reward as $r_{1,k} = LM_B(s_{mid,k})$, that indicates how it is natural in language B; then $s_{mid,k}$ is translated again in language A using the translation model $P(.|s;\Theta_{BA})$ and it is evaluated how this translation is consistent with the original sentence s. This estimate, that we name communication reward, is calculated as $r_{2,k} = \log P(s|s_{mid,k};\Theta_{BA})$. The total reward of $s_{mid,k}$ is $r_k = \alpha r_{1,k} + (1-\alpha)r_{2,k}$ where α is a hyperparameter. After all middle translation total rewards are calculated, we compute the gradient of the expected reward $E[r]$ with the respect to network parameters Θ_{AB} and Θ_{BA}, that is $\nabla_{\Theta_{AB}} E[r] = \frac{1}{K}\sum_{k=1}^{K}[r_k \nabla_{\Theta_{AB}} \log P(s_{mid,k}|s;\Theta_{AB})]$ and $\nabla_{\Theta_{BA}} E[r] = \frac{1}{K}\sum_{k=1}^{K}[(1-\alpha)\nabla_{\Theta_{BA}} \log P(a|s_{mid,k};\Theta_{AB})]$.

Finally, we update the two translation models. This process can be iterated many times until the convergence of the two translation models. In each round, one sentence is sampled from D_A and one from D_B, and we update the models according to the game beginning with the two sentences respectively.

4. Experimental Setup

Since the number of sentences that we have collected during the interaction between the human and the robot is still not sufficient to apply the Dual-NMT methodology, we focus on the experimental set-up and some first results related to Phase 1.

For our experiment, we decided to use a humanoid Nao robot of SoftBank Robotics. The choice fell on this kind of robot because it fits well with the implementation of soft sensors that form our somatosensory system: in fact, it is embedded with different sensors able to monitor many parameters, states, and events at the same time.

The Nao robot is embedded with different hardware base sensors that give information about the robot status. Among of these base sensors, in order to implement our soft sensors, we took into account the 25 actuators, the 4 switches located at the tip of each foot, the 9 touch sensors located on the head (front, rear, middle) and on the hands (back, left, right), the sonar located on the chest and some sensors linked to the battery.

For the experimental setup we implemented the following soft sensors whose detailed description can be found in [2]: *Current Pain*, *Temperature Pain*, *Caress Feeling*, *Bumper Pain*, *Exertion*, *Anxiety*. We sampled data from robot sensors with Choregraphe, a SoftBank Robotics desktop application that allows to monitor and control a Nao robot, making it perform some movements or say something. In particular, we used the Choregraphe Memory Watcher Panel to record our data from sensors and save the results as a CSV file. All data were sampled at 100Hz frequency.

Sampled data were then given as input to our soft sensors in order to generate a corpus of RoboLang sentences. This corpus was then annotated with some possible Italian translation. On these bilingual data, we experimented the approach of Phase 1.

4.1. From Soft Sensors to RoboLang

Once we sampled data from the above-mentioned sensors with Choregraphe, we split data into groups of 600 samples. Each group represents a RoboLang sentence.

Then we divided data of each sample based on the sensor type and gave them in input to the right soft sensor. We coded each non-zero soft sensor output value as a RoboLang word. As we said above a RoboLang word is a *(Roboception, Location, Value)* triplet, that indicates the occurrence of a roboception in a specific robot body part with a certain intensity.

Robot body parts were coded with a value K with $K = 0, 1, \cdots 19$ as follows: *(0 ⇒ Global), (1 ⇒ Head), (2 ⇒ RShoulder), (3 ⇒ LShoulder), (4 ⇒ RElbow), (5 ⇒ LElbow), (6 ⇒ RWirst), (7 ⇒ LWirst), (8 ⇒ RHand), (9 ⇒ LHand), (10 ⇒ Chest), (11 ⇒ RHip), (12 ⇒ LHip), (13 ⇒ RKnee), (14 ⇒ LKnee), (15 ⇒ RAnkle), (16 ⇒ LAnkle), (17 ⇒ RFoot), (18 ⇒ LFoot), (19 ⇒ Battery). Value* can assume a value in $\{VeryLow, Low, Normal, High, VeryHigh\}$ corresponding to a quantization of soft sensor output values. Each soft sensor outputs a real value in the $[0, 1]$ interval, which has been split in the following manner: $0 < output \leq 0.10 \Rightarrow$ *VeryLow*; $0.10 < output \leq 0.35 \Rightarrow$ *Low*; $0.35 < output \leq 0.65 \Rightarrow$ *Normal*; $0.65 < output \leq 0.90 \Rightarrow$ *High*; $0.90 < output \leq 1 \Rightarrow$ *VeryHigh*.

We generated a RoboLang corpus made up of 260 sentences. We have taken 176 of these and each of them has then been annotated with one or more possible translations in the Italian language. In the end, we obtained a bilingual corpus of 380 sentences. An example of Robolang-Italian sentences pair is:{*(Current Pain, 02, Normal), (Temperature Pain, 02, Normal)*} ⇒ {*Mi fa male la spalla destra (My right shoulder hurts)*}

4.2. Italian and RoboLang sentences sub-symbolic encoding

The implementation of both our encoding approaches uses a lookup table by exploiting two dictionaries: the "RoboLang-Italian" dictionary, that associates each RoboLang sentence with a list of possible translations in the Italian Language, and the "Italian-RoboLang" dictionary, that associates each Italian sentence with the corresponding RoboLang translation.

In the LSA-Word2Vec approach, we used the FastText Italian model [9] in order to encode the Italian sentences. For the RoboLang sentence encoding, we used the LSA apporach. In particular, at first we split and put in lowercase the sentences of our RoboLang corpus. From these, we obtained the term dictionary (composed of 137 distinct RoboLang terms, i.e., triplets) and consequently the RoboLang Sentence-Term matrix. Successively we trained our LSA model by using k=52 dimensions (or topics): therefore we built a RoboLang semantic space of dimensionality 52. The number of dimensions was chosen by using the coherence metrics [25]. In figure 2 we show the results of the experiments for the choice of the number of dimensions.

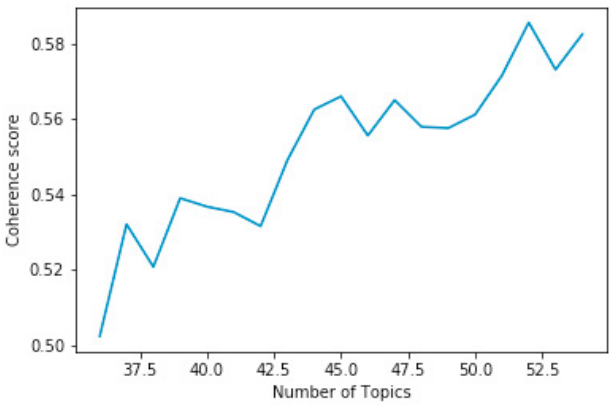

Fig. 2. Coherence values according to the number of topics. As we can see, the best coherence value is for k = 52.

In the pure Random Indexing approach, both Italian and RoboLang sentences are encoded with the RI technique. At first we split and put in lowercase our sentences. From these we generate the Italian vocabulary, made up of 43 words, and the RoboLang vocabulary, made up of 137 words. For each word we randomly generated an index vector of 2000 elements. We computed the context vectors by scanning our sentences with a 2-dimensional window. By using this approach, we emulate the condition that the robot does not know, a priori, the meaning of words.

The cosine function has been used as a semantic similarity measure for both the subsymbolic encodings.

4.3. Grounding procedure and vocabulary enhancement through interaction

The translation between RoboLang and Italian can be done in both directions. If we want to translate a sentence s from Italian to RoboLang, first we check if s is already in the Italian-RoboLang dictionary. In that case, we return the first of all the possible translations in RoboLang. Otherwise, we compute the similarity between the word embendings of s and all the sentences present in the Italian dataset, obtaining the sentence s_i with the best similarity score. In the LSA-Word2Vec approach, we use the FastText vectors; while in the Pure RI approach at first, we compute the index vectors of each new word and update the context vectors of s words, then we compute the encoding of s by averaging the context vectors of its words. If the score is greater than a threshold we add s to the dataset together with all the possible translations of s_i. Then we update the two dictionaries, and return the first one of these translations.

If the semantic similarity score is below the threshold, the sentence s cannot be directly translated, since its meaning is unknown to the robot. In this case, we let the robot asking the user to physically emulate the physical conditions that lead to the sensation described by s (e.g., by holding one of the robot hands, by caressing some of its body parts, and so on) so that a sampling procedure can start, and a RoboLang sentence can be associated with s. If this is not possible or if it is unsafe for the robot, the user can insert the RoboLang sentence by hand. Subsequently, we add s to the dataset along with the user RoboLang translation; we update the two dictionaries. If the RoboLang translation is not present into the dataset, we also update the doc-term matrix and the LSA model in the LSA-Word2Vec approach, and the context vectors in the Pure RI approach. The updating process of the whole model can be also run after a given number of sentences translations have been provided.

Similarly, if we want to translate a RoboLang sentence r to Italian, first we check if r is already present in the RoboLang-Italian dictionary. In that case, we return the first translation. Otherwise, in the LSA-Word2Vec approach we check if there is at least a new RoboLang word in r. If this is the case, the robot asks the user to give the Italian translation of r. The dictionaries are updated, and the whole LSA RoboLang model is computed. Otherwise, the embedding of the RoboLang sentence r is computed by using the projection in the LSA space. This makes it possible to search for the most similar RoboLang sentence r_i in the dataset. In the Pure RI approach, we first compute the index vectors of each new word and consequently update the context vectors of r words. Then we compute the encoding of r by averaging the context vectors of its words. If the similarity score is greater than a threshold we add r to the dataset with all the possible translations of r_i. We update dictionaries and the RoboLang Sentence-Term matrix and the LSA model for the LSA-Word2Vec approach, returning the first of these translations. Otherwise, we ask the user for an Italian translation of r, and we update both the dataset and the dictionaries accordingly, as well as the RoboLang Sentence-Term matrix and the LSA model or the index and context vectors.

4.4. Preliminary results

For the testing phase, we used two different corpora extracted for our bilingual corpus, one for each translation function. For the RoboLang - Italian translation testing, we used a corpus of 92 sentences. We obtained this corpus taking all the RoboLang sentences that were not used for the training corpus. For the Italian - RoboLang translation testing, we used a corpus of 162 sentences. We obtained this corpus by generating semantically similar sentences to the Italian sentences of the training corpus.

We tested both the approaches computing the accuracy of the two translation functions for different threshold values. The results are shown in figures 3 and 4. In both case, as the threshold value increases, the translation accuracy decreases. That occurs because with a certain threshold value the model cannot translate sentences that with a lower threshold value it was able to translate correctly. The LSA-Word2Vec approach performs very well with the RoboLang to Italian translation, with about the 96% of accuracy in the best case, while for the translation in the opposite direction the accuracy of the model is lower (about 62%) (Fig. 3). The accuracy of RI model is significantly worse (Fig. 4). As the threshold value increases, the accuracy of the model for the RoboLang - Italian translation decreases sharply from about 45% to less 10%. The accuracy for the Italian - RoboLang translation is stable because the sentences that model cannot translate due to the threshold value are sentences that it computed in a wrong way.

Figure 5 shows the comparison between the two approaches. We obtained these graphs by translating every sentence of the testing corpora, one by one, with both the approaches and computing the accuracy of each model until that sentence. As we can see, the LSA-Word2Vec model performs better than the Pure RI model in both translation directions.

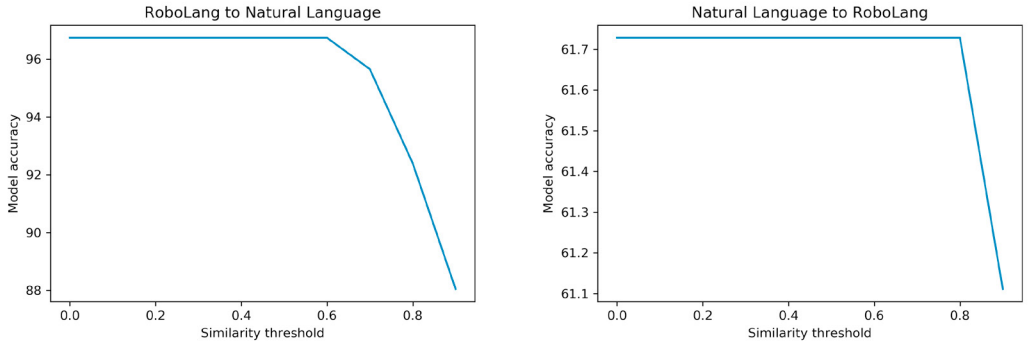

Fig. 3. Model accuracy for LSA-Word2Vec approach according to different threshold values.

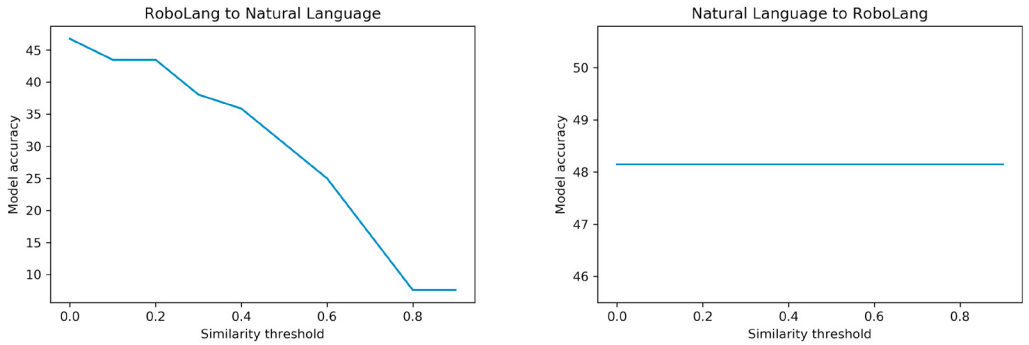

Fig. 4. Model accuracy for Pure RI approach according to different threshold values.

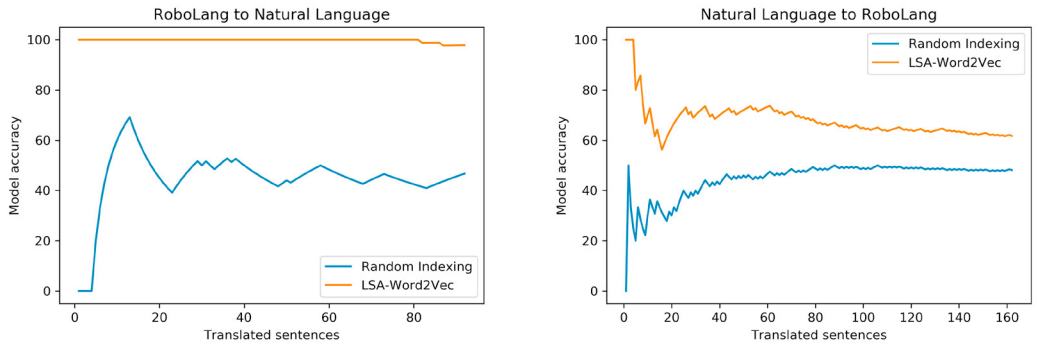

Fig. 5. Comparison between the two approaches.

5. Conclusions and Future Works

In this paper, we have presented the general idea of a system that could be capable to roughly emulate the learning process of babies with the interaction with their parent when they try to express their sensations and needs in natural language. We have implemented the whole framework, and a basic interaction-grounding has been provided. However, the parallel corpus that is being constructed is still meagre for being exploited even by the Dual-NMT approach. For the future, this will require a more extensive interaction and experimentation in order to let the whole approach work. We also consider as future aim, the possibility to trigger an artificial empathy in robots as proposed in [1], where in our case, the empathy could be triggered by words exchanged during the human-robot interaction.

Acknowledgements

This research has been partially supported by AMICO Project, CUP B46G18000390005; cod ARS01_00900 "Assistenza Medicale In COntextual awareness" decreto di concessione del 10 luglio 2018 prot. n.11598

References

[1] Anshar, M., 2017. Evolving robot empathy through the generation of artificial pain in an adaptive self-awareness framework for human-robot collaborative tasks. Ph.D. thesis.
[2] Augello, A., Infantino, I., Maniscalco, U., Pilato, G., Vella, F., 2017. The effects of soft somatosensory system on the execution of robotic tasks, in: Robotic Computing (IRC), IEEE International Conference on, IEEE. pp. 14–21.
[3] Augello, A., Infantino, I., Maniscalco, U., Pilato, G., Vella, F., 2018. Robot inner perception capability through a soft somatosensory system. International Journal of Semantic Computing 12, 59–87.
[4] Bloom, P., 2002. How children learn the meanings of words. MIT press.
[5] Broz, F., Nehaniv, C.L., Belpaeme, T., Bisio, A., Dautenhahn, K., Fadiga, L., Ferrauto, T., Fischer, K., Förster, F., Gigliotta, O., et al., 2014. The italk project: A developmental robotics approach to the study of individual, social, and linguistic learning. Topics in cognitive science 6, 534–544.
[6] Cangelosi, A., Ogata, T., 2016. Speech and language in humanoid robots. Humanoid Robotics: A Reference , 1–32.
[7] Chen, Y., Filliat, D., 2015. Cross-situational noun and adjective learning in an interactive scenario, in: 2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), IEEE. pp. 129–134.
[8] Di Gangi, M.A., Bosco, G.L., Pilato, G., 2019. Effectiveness of data-driven induction of semantic spaces and traditional classifiers for sarcasm detection. Natural Language Engineering 25, 257–285.
[9] FastText, 2019. FastText library for efficient text classification and representation learning. https://fasttext.cc/.
[10] Galipó, A., Infantino, I., Maniscalco, U., Gaglio, S., 2017. Artificial pleasure and pain antagonism mechanism in a social robot, in: International Conference on Intelligent Interactive Multimedia Systems and Services, Springer, Cham. pp. 181–189.
[11] He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.Y., Ma, W.Y., 2016. Dual learning for machine translation, in: Advances in Neural Information Processing Systems, pp. 820–828.
[12] Howell, S.R., Jankowicz, D., Becker, S., 2005. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. Journal of Memory and Language 53, 258–276.
[13] Lakoff, G., 1990. Women, fire, and dangerous things: what categories reveal about the mind. 1987. The University of Chicago Press, Chicago) Response Yes No Geometric Parameter Value Steering column angle (H18) 23, 390–550.
[14] Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104, 211.
[15] Landauer, T.K., Foltz, P.W., Laham, D., 1998. An introduction to latent semantic analysis. Discourse Processes 25, 259–284. doi:10.1080/01638539809545028, arXiv:https://doi.org/10.1080/01638539809545028.
[16] Langer, J., Bowerman, M., Levinson, S., 2001. The mosaic evolution of cognitive and linguistic ontogeny. Language acquisition and conceptual development , 19–44.
[17] Mangin, O., Filliat, D., Ten Bosch, L., Oudeyer, P.Y., 2015. Mca-nmf: Multimodal concept acquisition with non-negative matrix factorization. PloS one 10, e0140732.
[18] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .
[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.
[20] Morse, A.F., Cangelosi, A., 2017. Why are there developmental stages in language learning? a developmental robotics model of language development. Cognitive Science 41, 32–51.
[21] Nelson, R., 2001. The Somatosensory System: Deciphering the Brain's Own Body Image. volume 1 of *Frontiers in Neuroscience*. CRC Press.
[22] Nishihara, J., Nakamura, T., Nagai, T., 2017. Online algorithm for robots to learn object concepts and language model. IEEE Transactions on Cognitive and Developmental Systems 9, 255–268. doi:10.1109/TCDS.2016.2552579.
[23] Parisi, D., 2010. Robots with language. Frontiers in neurorobotics 4, 10.
[24] Pilato, G., Vassallo, G., 2015. TSVD as a statistical estimator in the latent semantic analysis paradigm. IEEE Transactions on Emerging Topics in Computing 3, 185–192.
[25] Röder, M., Both, A., Hinneburg, A., 2015. Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM international conference on Web search and data mining, ACM. pp. 399–408.
[26] Sahlgren, M., 2005. An introduction to random indexing .
[27] Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., Asoh, H., 2016. Symbol emergence in robotics: a survey. Advanced Robotics 30, 706–728.
[28] Vella, F., Pilato, G., Motisi, I., Gaglio, S., 2005. Automatic dictionary creation by sub-symbolic encoding of words, in: Neural Nets. Springer, pp. 113–119.