



Using Semantic Maps for Robust Natural Language Interaction with Robots

Emanuele Bastianelli¹, Danilo Croce², Roberto Basili², Daniele Nardi³

¹ DICII, ² DII - University of Rome Tor Vergata - Rome, Italy

³ DIAG - Sapienza University of Rome - Rome, Italy

bastianelli@ing.uniroma2.it, {croce,basili}@info.uniroma2.it
nardi@dis.uniroma1.it

Abstract

Modern robotic architectures are equipped with sensors enabling a deep analysis of the environment. In this work, we aim at demonstrating that such perceptual information (here modeled through semantic maps) can be effectively used to enhance the language understanding capabilities of the robot. A robust lexical mapping function based on the Distributional Semantics paradigm is here proposed as a basic model of grounding language towards the environment. We show that making such information available to the underlying language understanding algorithms improves the accuracy throughout the entire interpretation process.

1. Introduction

As robots are moving from industrial environments to consumer markets their human-like interaction capabilities with people is a key aspect. Human language is the most natural way of interaction for its expressiveness and flexibility. End-to-end communication processes in natural language are challenging for robots as for the deep interaction of cognitive abilities. Having a robot reacting to a command like “*Take the book on the table*” corresponds to a number of implicit assumptions. First, the environment must contain at least the entities *book* and *table*, and these must be perceived by the interlocutors; we also assume that the robot has an inner representation of the objects, e.g. through an explicit map. Second, lexical references to real world entities in utterances must be resolved through a *grounding* process [1] linking symbolic knowledge (e.g. words) to the corresponding perceptual information.

Notice that, while a large “natural” vocabulary is available to the user in order to express references to surrounding objects, situations or spatial relations, from a robotic perspective grounding particularly depends on both the quality of the robot’s lexicon and its ability to link linguistic symbols to its knowledge of the environment. Grounded reasoning seems required here to understand references towards the current state of the world, i.e. the grounded references that symbols exhibit towards objects of the environment, as well as their properties. Failures in such references are crucial. For example, assuming that no book is visually accessible while a table is known, no planning is possible for the robot: we would expect a failure in this sense to trigger a question like “*I don’t see any book in here. Can you help me?*”. Things are even more complicated as the robot may have no entry for *book*, so that such reference fails as for a partial lexical coverage. Finally, noise in the environment may lead to mis-transcriptions from the Automatic Speech Recognition (ASR) engine, so that the sentence may be misunderstood as “*Take the look on the label*”: this increases the relevance of linguistic robustness in the targeted Human-Robot Interactions (HRIs). The awareness that no *look* or *label*

exists in the room, would help in discarding mis-transcribed utterances or correcting them properly, by enforcing the suitable grounding (i.e. use *table* as target for the symbol *label* and *book* for *look*). The immediate consequence of the above observations is that robust HRIs strictly require flexible forms of grounding the linguistic symbols to objects. In addition, using this derived information early on during the spoken language interpretation process can be crucial. It follows that, while fully grounded understanding is the side effect of the entire Spoken Language Understanding (SLU) process, some hypothesis about the grounding of partial linguistic elements is highly helpful for improving the interpretation at different stages (e.g. recovering from ASR mistakes or tolerating lexical variability such as in *volume* as opposed to *book*). Several cognitive and psycholinguistic studies in fact have shown that a strict correlation exists between what we perceive and the interpretation we give to the natural language expressions we hear or read [2], [3], [4]. Our hypothesis is that the information coming from the perceptions of a robot can be used to enhance its spoken language understanding ability. Grounded information can thus be used to inject perceptual evidence in the interpretation process.

In this work, we propose a grounding function based on semantic information derived from Distributional Models (DMs) of the lexicon [5] as well as on phonetic similarity. DMs are broadly used in NLP to acquire semantic relations between words, e.g. quasi-synonymy, through the distributional analysis of large-scale corpora, in order to induce lexical rules such as “*volumes*” are semantically similar to “*books*”. Furthermore, the function also depends on the phonetic distance between words. References to the entities (and their symbolic names into the robot KBs) are thus mapped to distributions able to deal with uncertainty inherent to the grounding. In this way, the grounding capability of the robotic system is able to deal with unseen words as well as to tolerate possible mis-transcriptions coming from the ASR. Finally, we will show that this grounding evidence can improve the quality of the entire SLU chain.

In Robotics, grounding is often enabled by exploiting a symbolic representation for the robot perceptions. Variants are here *ad-hoc* built resources, e.g. handcrafted KBs containing properties of the represented objects, or the outcomes of automatic processes, e.g. visual object recognition or incremental compilation supported by the interaction with the user, as in *symbiotic autonomy* paradigm [6]. Recently, *semantic maps* [7] have been used to represent the perception of the world available to a robot. A semantic map is a knowledge base “*that contains, in addition to spatial information about the environment, assignments of mapped features to entities of known classes*”, i.e. a joint representation of perceptual and symbolic knowledge. Among the different properties about entities, a semantic map usually contains linguistic references to them, specified when maps are built with the human assistance, e.g. by point-

ing and naming the entities in the environment, following the so-called *Human Augmented Mapping* [8] paradigm [9], [10], [11]. In this sense, a map containing entities, such as a `book`, characterizes also their linguistic references, e.g. the name `book` for which `hasName(book, book)` is true. The problem of grounding natural language has been tackled in Robotics by several works, most of them suggesting some form of supervision to learn the mapping between words and the robot perception of the world. Some of them relied on the multi-modal “show-and-tell” technique, with the user showing an object, while naming it [12], [13]. In other works such experience is surrogated by artificial training examples, and the mapping is learned through a classification scheme. Statistical methods are applied to learn the linking between words and visually perceived objects [14], [15], or to map a command into a parse structure representing the grounding of the entire sentence [16], [17] and [18] in terms of actions, spatial relations and entities.

In the rest of the paper, Section 2 discusses a lexicalized approach to grounding that has been used within an extension of the architecture for SLU for HRI as the one introduced in [19] (Section 3). An original contribution of this work is thus to provide a more robust way of linking words to entities without the need of an explicit categorization, but adopting a linguistic consensus approach, i.e. exploiting lexical usages across large-scale corpora. Grounding is here not restricted to a dictionary available in a (possible small) set of examples, but share properties naturally available to the human community of a natural language. Second, no previous approach in Robotics injects grounded information in the interpretation process. Even though some works jointly use the linguistic and perceptual information, the resulting interpretation process is restricted to a known (i.e. seen in the training phase) environment. No improvement is available against scene changes or unseen words and none of the previous approaches deals with mis-transcription errors. These contributions are supported by evaluations as reported in Section 4.

2. Grounding According to Lexical Competence

A general assumption of our work is that grounding corresponds to map entities mentioned in a command against objects of a semantic map, where the latter are located and have an associated name, provided through a Human Augmented Mapping (HAM) process, and specified by the `hasName(·, ·)` property. For example, when the map contains an entity `e`, e.g. a `book`, then a property `hasName(e, we)` such as `hasName(book, book)`, can be associated with it, as the word `we`, i.e. `book`, represents the `book` name.

Given a word `w` occurring in a command `c` used to refer to an entity in the environment, the function expresses a likelihood distribution over all entities `e` contained in the semantic map. The grounding function corresponds to a fuzzy membership, i.e. a confidence score, to each possible grounding $\langle w, e \rangle$ to entities `e` corresponding to names `we`. It measures the confidence that `w` can be grounded to entity `e` and depends on two major evidences: the **phonetic similarity** between word `w` and the lexical reference for `e`, i.e. `we` (e.g. “`look`” vs. “`book`” wrt `book`); the **lexical similarity** that measures a paradigmatic relatedness, e.g. the quasi synonymy, between a `w` and the lexical reference `we`, e.g. “`volume`” and `book`). Notice that **phonetic and lexical similarity** usually act together, as mis-transcriptions, e.g. “`valium`”, may well correspond to quasi synonyms, e.g. “`volume`”, for a `book`. Lexical similarity could be surrogated by a

priori handcrafting the lexicon containing all admissible alternatives for a given `we`. Although this solution can be applicable in small or controlled scenarios, it is very expensive, whereas existing lexicons, e.g. WordNet [20], are not easily generalizable across languages and domains.

In order to characterize the words admissible for a target `we`, we refer to a Distributional Model (DM) of the lexicon [5]. DMs are intended to acquire semantic relationships between words in an unsupervised fashion, mainly by looking at the words usage in large scale corpora. The foundation for these models is the *Distributional Hypothesis* [21], i.e. words that are used and occur in the same contexts tend to share similar meanings. In recent years, DMs have been at the basis of many advances in NLP, and different methods have been proposed to derive them in efficient ways. Main approaches estimate semantic relationships in terms of vector similarity. As an example, a word `wk` is represented through a vector \mathbf{w}_k whose dimensions encode all words co-occurring with `wk`. As discussed in [22], words in *paradigmatic* relations, e.g. quasi-synonymy tend to co-occur with the same words and they will be represented through similar vectors. The word representations fostered in this paper are obtained by applying a Recursive Neural Network architecture [23]. This approach allows to derive a projection function $\Phi(\cdot)$ of words into a geometrical space, i.e. the vector representation for a word `wk` $\in \mathbb{W}$ is obtained through $\mathbf{w}_k = \Phi(w_k)$. In such embedded space, words with similar meaning are represented similarly so that the distance function between vectors reflects the linguistic relatedness. Given two words `w1` and `w2`, their similarity relatedness *sim* is implicitly defined and can be estimated as the cosine similarity between the corresponding vectors $\mathbf{w}_1, \mathbf{w}_2$ in the space, i.e. $sim(w_1, w_2) = \frac{\mathbf{w}_1 \cdot \mathbf{w}_2}{\|\mathbf{w}_1\| \|\mathbf{w}_2\|}$. For example, following such approach, given the word “`book`”, the closest words in the space, i.e. the most related words, are “`booklet`”, “`volume`” “`chapbook`”, “`guidebook`” and “`manuscript`”.

Formally, we define $S_\tau^e = \{w \mid sim(w, w_e) \geq \tau\}$ as the set of all the words having a semantic relation with `we`, derived by the DM through the *sim* function. The threshold τ allows to restrict the set only to the most related words. Given a word `wc` in the transcribed vocal command that references an entity in the environment, the function find the entity `e` whose `we` maximizes the similarity with `wc`. Moreover, as mis-transcriptions may occur, we allow to ground different but phonetically similar words, e.g. “`valium`” vs. “`volume`”, penalizing this grounding proportionally to the phonetic distance. Given a word `wc` the final grounding function value will be computed as:

$$g(w_c, e) = \max_{w \in S_\tau^e} (ph(w_c, w) * sim(w_e, w)) \quad (1)$$

where *ph* is the normalized phonetic similarity¹ between two words, evaluated as the inverse of the Levenshtein distance [24] on their transcriptions. More refined methods can be used for this factor, and they will be investigated in further works. By evaluating the grounding function against all the entities `e` in the semantic map, we obtain a likelihood distribution of the possible groundings with respect to a pronounced word `wc`.

Let us consider a map with a book `b1` and a table `t1`, with associated names `wb1=book` and `wt1=table` respectively. Consider a command containing the word `wc` “`volume`” to be grounded. In order to evaluate the confidence in grounding w.r.t. `b1`, $g(\text{“volume”}, b1)$ is computed: the *sim* function retrieves the set $S^{b1} = \{\text{“book”}, \text{“booklet”}, \text{“volume”}, \dots\}$. For each

¹As the cosine similarity hold in [0,1] the phonetic similarity is normalized and $ph = \frac{1}{(dist+1)}$, where *dist* is the Levenshtein distance

$w \in S^{b1}$, phonetic similarity against “volume” is multiplied by the value of $sim(w_{b1}, w)$. With a $sim(b1, “volume”)=0.78$ and a phonetic similarity $ph(“volume”, “volume”)=1$, the resulting $g(“volume”, b1)=0.78$. Then, $g(“volume”, t1)$ is evaluated with $S^{t1} = \{ “table”, “desk”, \dots \}$. The value of $g(“volume”, table)$ is lower than the value of $g(“volume”, b1)$, as the phonetic similarity between w_c and every word in S^{t1} penalizes the score of $g(“volume”, t1)$. Hence, w_c is grounded in $b1$.

3. Perception-supported SL Understanding

In order to show the effects of the information derived from the semantic maps, we applied it to an existing system for SLU for HRI [19]. The SLU processing chain is a cascade of components based on Machine Learning methods whose work-flow is hereafter shortly summarized. Given an audio input captured with a microphone, an ASR engine, based on the Google API², provides a n -best list of transcription hypotheses. Each transcription hypothesis is then lemmatized, POS-tagged and parsed through the Stanford CoreNLP³. A speech re-ranking module [25] produces a new ranked n -best list, promoting promising hypotheses in higher positions. The confidence in a hypothesis is computed according to a model of syntactic consistency and lexical semantic coherence trained over the typical structures of spoken commands. Support Vector Machines (SVMs) are used to this end, by exploiting a combination of several linguistic Kernels functions. The first hypothesis of the new rank is then analyzed by semantic parser discussed in [19] and provides information about actions and their arguments expressed in a sentence. The representation formalism is derived from Frame Semantics [26] and provides a linguistic and cognitive basis to the interpretation. According to this linguistic theory, actions or, more generally events, are modeled as *semantic frames*. These are micro-theories about real world situations, e.g. the action of TAKING, specifying the set of participating entities, called *frame elements*, e.g. the THEME, representing the object taken during the aforementioned action. For example, for the sentence “take the book”, the corresponding parsing is $[take]_{TAKING} [the\ book]_{THEME}$. The overall parsing process exploits a cascade of SVM classifiers.

In our perspective, an instantiated frame correspond to a robot plan, and frame elements to plan arguments, so that the semantic parsing process represents a first step towards action grounding. The second step consists in the grounding of the expressed entities in the real world, so mapping words w_c mentioned in frame elements to entities e in the semantic map. While in [19] this step corresponds to the retrieval of entities e named by w_c , i.e. only $w_c = w_e$ is considered, we generalize here such mechanism through Equation 1. It must be noticed that lexical information contained in semantic maps is also beneficial to the ASR stage. We here will also verify that grounding hypothesis may be exploited to select the right hypothesis in the ASR n -best list. ASR engines use to provide multiple solutions to a given audio input. Words with similar spellings can be easily mis-recognized, especially when the background noise is significant. Sometimes, the right solution may not be provided at all. Our aim is to use confidence scores given by the grounding function 1 to support the selection of hypotheses containing more likely grounded references than others. This would help the re-ranking module to better retrieve hypotheses with direct references to the current envi-

²www.google.com/intl/it/chrome/demos/speech.html

³nlp.stanford.edu/software/corenlp.shtml

ronment. Moreover, the approach would assure that, at least, hypotheses containing mis-transcriptions that are better resembling the entities in the map are ranked in a higher position. As the re-ranking module exploits SVM based algorithms, we can make available this grounding information through four additional features, than the one discussed [25], corresponding to the *Mean* and *Standard Deviation* scores across all references w_c in the sentence, as well as, the *Lowest* ($\min_{c,e}(g(w_c, e))$) and *Highest* ($\max_{c,e}(g(w_c, e))$) grounding scores.

4. Experimental Evaluation

In our experimentation, we aim first at measuring whether our grounding function improves the coverage against targeted entities as discussed in Section 4.1. In Section 4.2, we test the contribution of our grounding hypothesis, derived from the semantic map, when injected in the ASR re-ranking phase. In Section 4.3 both mechanisms are applied to the entire SLU chain.

Commands used during the experiments belong to HuRIC [27], a corpus of spoken commands for robots in the house servicing scenario⁴, paired with their correct transcriptions. All transcriptions are tagged at different linguistic levels, from morphology, syntax to different semantic formalisms. Commands are grouped in three datasets: the Grammar Generated (GG), the Speaky for Robots (S4R) and the Robocup (RC). These datasets are characterized by an increasing complexity: commands in the GG are simple, as “go to the kitchen”, while colloquial and courtesy forms are used in the RC, as “could you please turn on the lamp”. Even though commands in HuRIC are representative of possible interactions, they do not contain enough lexical variations against the underlying maps. In order to realistically account for these phenomena, we build different “lexical” variants of the semantic maps corresponding to the HuRIC commands. First the set of entities E including all entities referenced in any considered command is collected. Then, for each $e \in E$, possible lexical alternatives w_e as possible names of e ’s is added by sampling words from the DM (described in Section 2) and WordNet [20] dictionaries. Every association between an entity e and a name w_e has been validated by three annotators. For each command, 10 semantic maps are randomly populated by selecting one w_e for each $e \in E$, and adding also the property $hasName(e, w_e)$. The DM used within the *sim* function has been acquired according to a Skip-gram model [23] through the word2vec tool⁵: we derived 250 dimensional word vectors for more than 110,000 words, by using the UkWaC corpus⁶ a large scale web-pages collection made of 2-billion words.

4.1. Evaluating the Grounding function

First, we evaluate the accuracy of our grounding function, against maps and annotations, relying only on the correct transcriptions. Grounding accuracy is measured as the percentage of entities correctly retrieved from a targeted semantic map referenced by a command. The maps generated from the commands contained 28, 19 and 29 objects and locations for the GG, S4R and RC dataset respectively. In our setting, a semantic map may contain more than one grounding candidate for each word w_c , as multiple entities of the same type may be present, whose names overlaps, e.g. two books both named as *book*. Thus, given a semantic map and the set E of entities contained in it, different grounding functions that select a subset $\Gamma \subseteq E$ of

⁴Available at www.http://sag.art.uniroma2.it/huric

⁵word2vec is available at code.google.com/p/word2vec/. The settings are: *min-count=50*, *window=5*, *iter=10* and *negative=10*.

⁶<http://wacky.sslmit.unibo.it/doku.php?id=corpora>

the candidate entities are compared. The *Identity* ($=$) function, already used in [19], selects entities whose names correspond to the input w_c , i.e. $\Gamma = \{e | w_c = w_e\}$; the *Argmax* (Argmax) function returns only the entity e that maximizes the grounding function, i.e. $\Gamma = \{e | \operatorname{argmax}_{e \in E} (g(w_c, e))\}$, with $g(\cdot, \cdot)$ defined in Eq. 1. When the function *Single Threshold* (θ) is used, all groundings for which $g(\cdot, \cdot)$ is higher than a threshold θ are returned, i.e. $\Gamma = \{e | g(w_c, e) > \theta\}$, where θ is estimated on a development set. Finally, in *Mean Value* ($\mu + n \cdot \nu$) we consider valid all the entities whose grounding score is higher than the sum of the mean of the grounding scores over $e \in E$ with respect to w_c , and its standard deviation σ_{w_c} , i.e. $\Gamma = \{e | g(w_c, e) > \mu_{w_c} + n \cdot \sigma_{w_c}\}$, where the factor n increases flexibility for $n \geq 0$ over distributions $g(\cdot, \cdot)$. For each dataset, all the contained commands are grounded on the corresponding 10 variations of the semantic map, by applying the above functions. Precision (P), i.e. the percentage of correctly grounded entities among the retrieved ones, and Recall (R), i.e. the percentage of correct grounded entities among the set of *gold* ones, give rise to the traditional F-Measure, as their harmonic mean. Table 1 reports these performance metrics for the different functions over the datasets in HuRIC. The Identity function is the most strict, thus achieving the highest Precision, but very poor Recall. DM-oriented grounding functions clearly improve the coverage and their results strongly depend on the acceptance threshold: lower values (e.g. $n = 0$) correspond to higher recall. A good compromise is achieved with the θ function although its estimation requires a training set. Finally, the $\mu + 3\sigma$ policy achieves the best results.

Table 1: Scores of the Grounding Functions.

	GG			S4R			RC		
	P	R	FI	P	R	FI	P	R	FI
$=$	100.0	31.2	47.6	100.0	34.4	51.2	100.0	37.8	54.9
Argmax	58.1	58.1	58.1	64.3	66.2	65.2	49.0	71.2	58.5
θ	60.5	69.8	64.8	46.9	75.4	57.8	62.7	52.5	57.2
μ	11.2	87.4	19.8	19.8	88.2	32.4	8.5	94.2	15.6
$\mu + \sigma$	30.1	77.0	43.3	34.4	78.4	47.8	20.0	90.7	32.8
$\mu + 2\sigma$	53.2	72.1	61.2	60.4	69.0	64.4	41.2	83.2	55.1
$\mu + 3\sigma$	73.0	58.9	65.2	81.4	56.5	66.7	63.1	67.7	65.3

4.2. Using grounding evidence to improve ASR accuracy

In this section we discuss the results obtained by applying a re-ranking stage fed with the information derived by the semantic map (Re-rank + SM) and comparing it with simple re-ranking (Re-rank) or no re-ranking (i.e. only the transcription provided by Google is returned) (No Re-rank). A 4-fold evaluation strategy is applied, with one fold for testing. System performances are measured in terms of Precision at 1 (or P@1), i.e. the percentage of correctly transcribed sentences that occupy the first position in the rank. Each command has been randomly paired with one of the 10 semantic maps, since the features derived from the grounding function depend on a specific map. Table 2 shows the mean and the standard deviation (across the 4 folds) of the performances for each dataset in HuRIC, and the Word-Error Rates (WER). In two datasets over three, using only linguistic information is beneficial to the re-ranking process (GG and S4R). No significant improvement is observed over the RC dataset, characterized by an higher linguistic complexity. Features derived from the semantic map, are beneficial in two cases over three. While the performance on the S4R remains unchanged, the scores on both the GG and RC increase from 83.4 to 94.4 and from 78.9 to 79.1, respectively. Although the difference seems not significant, it is worth noticing that results are compared with a state-of-the-art ASR system, i.e. Google. Moreover, standard deviation always decreases, suggesting that the re-ranker stabilizes and performance drop risks are lowered.

Table 2: Impact on the Speech re-ranking. P@1 and WER

	No re-rank			Re-rank			Re-rank + SM		
	P@1	WER		P@1	WER		P@1	WER	
GG	83.5	± 11.5	2.75	92.4	± 5.4	1.09	94.4	± 3.3	0.8
S4R	85.5	± 6.7	3.96	91.3	± 2.8	1.49	91.3	± 2.8	1.49
RC	78.9	± 6.7	3.85	77.9	± 8.4	3.30	79.1	± 6.0	3.47

4.3. Evaluating the entire processing chain

We tested the whole HRI processing chain enhanced with the re-ranking augmented with features extracted from the semantic map, starting from audio sources and outputting a completely grounded command. We compared two configurations: one that makes use of the identity function as a simple grounding mechanism ($=$), and other that exploits the $\mu + 3\sigma$ grounding function. The models used for re-ranking are the ones that produced the best results in Section 4.2. For the testing, a leave-1-out strategy is applied. For each command, the semantic parser is trained over all the other sentences, and the grounding function is applied to its output on the test sentence. Table 3 reports the reachable accuracy of the SLU chain as the rate of correctly grounded entities, as in Section 4.1. Over the three datasets, the grounding performance increases from 24.4% to 38% for the GG, from 39.0% to 40.2% for the S4R and from 49.7% to 55.7% for the RC, with a relative improvement of 55.5% and 12.1% on the GG and RC respectively. The improvement on S4R is not significant as for the trade-off between gain in recall and drop in precision: this is possibly due to the $\mu + 3\sigma$ function that returns too many candidates.

Beware that, although the above may look not striking, they realistically estimate the overall complexity of the processing chain. As demonstrated in [28], the overall upper bound for semantic parsing chains, as evaluated in not restricted texts, has an estimated accuracy of 70%. This factor, as for the interaction of several chain modules, drops down to a rough F1 estimate of $.79 \times .7 \times .65 = .36$ for the RC dataset, respectively for the re-ranking (0.79), parsing (0.7) and grounding (0.65). Moreover, our SLU chain is designed to enforce flexibility in the NL communication, hence balancing precision and recall.

Table 3: Performances of the whole processing chain

	GG			S4R			RC		
	P	R	F	P	R	F	P	R	F
$=$	51.4	16.0	24.4	75.5	26.2	39.0	81.8	35.6	49.7
$\mu + 3\sigma$	42.0	34.6	38.0	46.4	35.5	40.2	54.6	56.9	55.7

5. Conclusions & Future Work

In this work, perceptual information is shown effective in enhancing the language understanding capabilities of a robot. A function to ground nominal references as found in robotic commands on lexically augmented semantic maps is introduced. Such function robustly models lexical preferences with respect to potential mis-transcriptions of an ASR engine and variable linguistic expressions of the commands. Furthermore, according to the thesis that available perceptual information is beneficial to improve the interpretation process of spoken commands, we investigated the impact of grounding evidence on a concrete SLU process for HRI. Results show that the selection of the best ASR hypothesis can be improved (up to $\sim 55\%$) when ASR re-ranking informed about grounding is applied.

Further work will investigate a wider use of the information contained in semantic maps, for example taking into account also spatial relations among entities. Robots in fact are provided with perception systems of growing complexity: injecting this information within a statistical NL learning setting is promising and effective. Finally, grounding based on more accurate measures of phonetic similarity should be investigated.

6. References

- [1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.
- [2] R. M. Cooper, "The control of eye fixation by the meaning of spoken language : A new methodology for the real-time investigation of speech perception, memory, and language processing," *Cognitive Psychology*, vol. 6, no. 1, pp. 84–107, Jan. 1974.
- [3] M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy, "Integration of visual and linguistic information during spoken language comprehension," *Science*, vol. 268, pp. 1632–1634, 1995.
- [4] F. Huettig, J. Rommers, and A. S. Meyer, "Using the visual world paradigm to study language processing: A review and critical evaluation," *Acta Psychologica*, vol. 137, no. 2, pp. 151–171, Jun. 2011.
- [5] H. Schütze, "Word space," in *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, 1993, pp. 895–902.
- [6] S. Rosenthal, J. Biswas, and M. Veloso, "An effective personal mobile robot agent through symbiotic human-robot interaction," in *International Conference on Autonomous Agents and Multiagent Systems*, vol. 1, May 2010, pp. 915–922.
- [7] A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 915–926, 2008.
- [8] E. A. Topp, "Human-Robot Interaction and Mapping with a Service Robot: Human Augmented Mapping," Ph.D. dissertation, Royal Institute of Technology, School of Computer Science and Communication, 2008.
- [9] G. Kruijff, H. Zender, P. Jensfelt, and H. Christensen, "Clarification dialogues in human-augmented mapping," in *Proceedings of the 1st Annual Conference on Human-Robot Interaction*, 2006.
- [10] S. Hemachandra, T. Kollar, N. Roy, and S. Teller, "Following and interpreting narrated guided tours," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Shanghai, China, 2011, pp. 2574–2579.
- [11] E. Bastianelli, D. D. Bloisi, R. Capobianco, F. Cossu, G. Gemignani, L. Iocchi, and D. Nardi, "On-line semantic mapping," in *16th International Conference on Advanced Robotics*, Nov 2013.
- [12] D. Roy, K. yuh Hsiao, P. Gorniak, and N. Mukherjee, "Grounding natural spoken language semantics in visual perception and motor control," AAI, Tech. Rep., 2002.
- [13] J. Connell, E. Marcheret, S. Pankanti, M. Kudoh, and R. Nishiyama, "An extensible language interface for robot manipulation," in *Artificial General Intelligence*. Springer, 2012, pp. 21–30.
- [14] S. Guadarrama, L. Riano, D. Golland, D. Gohring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding spatial relations for human-robot interaction," in *The IEEE/RSJ International Conference on Intelligent Robots and Systems*, November 2013.
- [15] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using Ilda," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3943–3948.
- [16] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Proceedings of the 5th Annual Conference on Human-Robot Interaction*, Piscataway, NJ, USA, 2010, pp. 259–266.
- [17] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Approaching the symbol grounding problem with probabilistic graphical models," *AI Magazine*, vol. 32, no. 4, 2011.
- [18] J. Fasola and M. J. Matarić, "Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Tokyo, Japan, 2013.
- [19] E. Bastianelli, G. Castellucci, D. Croce, R. Basili, and D. Nardi, "Effective and robust natural language understanding for human-robot interaction," in *Proceedings of 21st European Conference on Artificial Intelligence 2014*. IOS Press, 2014.
- [20] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [21] Z. Harris, "Distributional structure," in *The Philosophy of Linguistics*, J. J. Katz and J. A. Fodor, Eds. Oxford University Press, 1964.
- [22] M. Sahlgren, "The word-space model," Ph.D. dissertation, Stockholm University, 2006.
- [23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.
- [24] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.
- [25] R. Basili, E. Bastianelli, G. Castellucci, D. Nardi, and V. Perera, "Kernel-based discriminative re-ranking for spoken command understanding in HRI," in *Proceedings of the XIII Conference of the Italian Association for Artificial Intelligence*, vol. 8249. Springer, 2013, pp. 169–180.
- [26] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *Proceedings of ACL and COLING*, 1998, pp. 86–90.
- [27] E. Bastianelli, G. Castellucci, D. Croce, R. Basili, and D. Nardi, "HuRIC: a human robot interaction corpus," in *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, may 2014.
- [28] R. Johansson and P. Nugues, "The effect of syntactic representation on semantic role labeling," in *Proceedings of 22nd International Conference on Computational Linguistics 2008*, Stroudsburg, PA, USA, 2008, pp. 393–400.