Spatial Language Processing for Assistive Robots with "Deep" Chunking and Semantic Grammars

Tatiana Alexenko
University of Missouri—Columbia
ta7cf@mail.missouri.edu

Marjorie Skubic
University of Missouri—Columbia
skubicm@missouri.edu

Zhiyu Huo
University of Missouri—Columbia
zhiyuhuo@mail.missouri.edu

Abstract

This paper presents a semantic spatial language grammar and a novel chunking method that allows nested structures to be encoded as a single label. The proposed semantic grammar, when used with a cognitive architecture described elsewhere, makes it possible for a mobile robot to follow complex, human-generated spatial descriptions for a fetch task. The semantic grammar is based on an interdisciplinary analysis of a corpus of human generated indoor spatial language. The "deep" chunking method facilitates encoding deep grammatical structures into a single-level label. The proposed method has been successfully used by an autonomous agent in a virtual environment as well as by a physical mobile robot. The deep chunking approach allows fast, feature-based machine learning methods usually used for shallow chunking to be used for deep nested parsing which better supports real-time interaction with a robot. Preliminary accuracy results are presented along with possible improvements and additional applications.

Introduction

Human comprehension of spatial language is a complex activity. Consider the comprehension of: "Your eyeglasses are behind the lamp on the table in the room on your left". First, the spatial term behind is a qualitative term rather than a precise metric location. Second, this region depends upon the interpretation of the spatial term, which could be based on the perspective of the speaker or addressee, the orientation of the room, or the axes of the reference objects (lamp, table). Third, the description is typically understood within a conversational context that includes speaker's assumptions about the addressee's capabilities and knowledge (Clark et al. 1983) and the establishment of a common ground (Clark 1996). Despite these complexities, a human would be able to follow these instructions with ease. In sharp contrast, the comprehension of such spatial descriptions is particularly problematic for robots.

This paper builds on top of the Human Robot Interaction (HRI) architectures proposed in Skubic and Carlson (2011-2013). This architecture includes an intelligent mobile agent, either physical or simulated in a virtual environment, designed to navigate an indoor domestic setting using directions to a target item given by a human. Skubic and Carlson (2012) referred to this as the "fetch" task. In contrast to mobile agents in other works (Matuszek 2010, Levit 2007) concerned with indoor spatial navigation, this agent has enhanced sensing capabilities. Using the Microsoft Kinect color and depth cameras, it can recognize furniture, walls and target objects and use the Histogram of Forces (Skubic 2003) to resolve basic spatial relationships. While these capabilities simplify the task of mapping from language to action, they do not solve the problem completely. The methods proposed in this paper aim to utilize these capabilities to facilitate a more informed search of an indoor space for a target object.

Our prior work (Skubic and Carlson 2011-2013) includes extensive linguistic analysis of a corpus of real spatial language collected from older and younger adults in which linguistic difference were observed across age and addressee groups. This analysis served as a basis for the proposed semantic spatial language grammar. The spatial grammar proposed in this paper was used successfully by a mobile robot in a virtual environment in Skubic and Huo (2013). The purpose of this paper is to discuss the basis for the semantic grammar and the necessary Natural Language Processing (NLP) methods in more detail. Preliminary results for automatic chunking and Part-of-Speech tagging are presented. Some additional uses for the grammar, such as generation of more training data and linguistic summaries are also discussed.

Indoor Spatial Language Corpus

Carlson and Skubic. et. al. (2011) conducted a human subject experiment that resulted in the creation of an indoor spatial language corpus (CSISL) used in this paper. The experiment was conducted in a virtual environment shown in Figure 1 and the participants from younger (college students) and older (64+ years old) age groups were asked to tell a robot or human avatar where a target object is or how to get to it. The oral responses were recorded and later manually transcribed. A total of 1024 spatial descriptions were collected: half from younger adults and half from older. Similarly, half of each subgroup was given "how do you get to an object?" prompts and half "where is the object?" prompts. Another variable of the experiment was the addressee type: human or robot avatar.

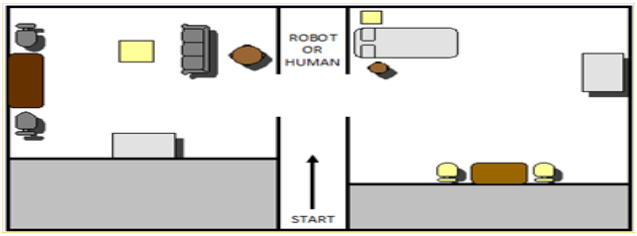

Figure 1. The overhead map of the environment.

Static "Where" and Dynamic "How" Descriptions

In contrast to related work that exclusively focused on dynamic "how" descriptions (Levit 2007, Tellex 2011, Matuszek 2010), the Carlson and Skubic et. al. (2011) experiment also captured static "where" descriptions by asking a subgroup of the participants "where" the target object is. This manipulation was included to investigate different types of spatial language that is offered naturally for the fetch task. The experiment yielded a variety of spatial language, including hybrid static/dynamic combinations. Figure 2 shows sample descriptions from the corpus to give the reader a better idea of what is meant by static "where" and dynamic "how" descriptions. The "where" prompt did not always yield static descriptions, and the "how" prompt did not always yield dynamic descriptions. These natural language descriptions pose challenges for translation into robot commands, because directions are often not sequential and contain a variety of directional terms.

Where:
The wallet is in the room on your right around the bed and on the bedside table

How:
Go to the room on your right…go past the couch… behind the couch there's an end table… the mug is on the end table

Hybrid:
The notepad is in the room on your right walk in and it's on the white...dresser to your left next to an empty box of kleenex.

Figure 2. Sample descriptions from the corpus. The last has both static and dynamic components.

Furthermore, "where" descriptions, by omitting ambiguous directional terms such as "left" and "right" or providing other indications of location, such as additional details about furniture and other landmarks, eliminate the issues that arise when there is perspective ambiguity. In (Carlson 2013) it was shown that "where" descriptions from this spatial language corpus were easier for humans to follow; in a separate experiment where human subjects were asked to follow descriptions from the corpus, participants were 35% more likely to correctly select the table where the target object was located when following "where" descriptions than "how" descriptions.

Template Corpus

In addition to the primary corpus of 1024 actual utterances (CSISL), Carlson et. al. (2011) also created a template corpus consisting of 149 descriptions based on templates derived from the analysis of the primary corpus. Templates were created to capture language structure and word frequency that were common to the spatial language descriptions logged for each manipulation: how vs. where instructions, robot vs. human addressee, and older adult vs. younger adult subjects. Due to some repetition across

How + Human Addr. + Younger Adult Template
[Go/Move/Take/Walk] + [straight/forward] + [turn/look/take] + [left/right] + [path landmark]
Original Example:
Take a couple steps forward. Take a left through the open door. Go to the foot of the bed and the cell phone is on the table.
Built From This template:
Go forward and turn right and the cellphone will be on the table to the right in front of the couch.

Where + Human Addr. + Younger Adult Template
[Object/It] + [Is In] + [room] + [on/to] + [left/right] + [additional direction] + [path landmark]
Original Example:
The wallet is in the room to Brian's left in the corner on a nightstand next to the bed.
Built From This template:
The wallet is in the room to the left on the table against the left wall.

Figure 3. Examples of a template for a particular subgroup, a description from the CSISL corpus matching this template and a description generated from this template.

different addressee, age and "how" or "where" manipulations, there were only 149 unique templates. These templates and the analysis they were based on influenced the semantic grammar proposed in this paper. Figure 3 provides an example of a template and corresponding original and generated template description. The templates were created manually from words most frequently used by the experimental subgroups.

Semantic Spatial Grammar

The templates discussed above and the analysis of the linguistic differences used by experimental subjects motivated the development of the spatial language grammar presented in this paper.

Semantic Part-of-Speech Tags

Skubic et. al. (2012) observed that experimental subjects used furniture very often in the descriptions they gave. This was true for all of the groups. Another commonly used landmark type was "house structure" such as wall, opening and door. The word "room" and the variations such as "bedroom" were also common. All of these words happen to be nouns as are the target objects and other nouns in the descriptions. To distinguish these different types of nouns, additional semantic Part-of-Speech (PoS) tags were defined: FUR for furniture, STR for house structures and RM for the word room or variations such as "bedroom". Additionally, the preposition "on" was given the "ON" tag and interjections such as "uhh" and "ok" were labeled "IGN" which stands for "ignore". Aside from these changes and additions, the Penn Treebank (Marcus 1993) PoS tag set was used.

The use of these semantic tags is meant to simplify aspects of chunking and further steps such as perception. As discussed in Skubic (2012), different visual object recognition approaches can be used for large vs. small objects, and separating furniture from the smaller target objects becomes useful. Walls and doors will require different perception approaches than furniture and objects—their locations can come from a preloaded or learned map, for example. The semantic PoS tags also assist with chunking by providing a more precise feature to the chunk label classifier than just a "noun" label. This is further discussed in the next section.

Semantic Grammar and Nested Chunks

Table 1 lists the semantic chunk types and their abbreviations and Figure 4 shows the proposed grammar applied to a real description. The semantic grammar proposed in this paper is specific to the "fetch" task described in (Skubic 2012) and the CSISL corpus. However, a large part of it can be used for more general indoor spatial language processing. For example, as shown in Figure 1, the spatial language description starts in the hallway because that is where the speaker was located when prompted by the experimenter. This motivated the separation of the spatial descriptions into two parts: outside of the target room and inside of the room. Although this is a separation motivated by the task and the corpus, it's likely that many indoor spatial language tasks will involve navigating between rooms. The next separation deals with the difference between parts of the description that describe a goal state or goal object vs. parts that describe what needs to be done or observed to reach the goal state.

Outside Room Target Phrase	ORMTP
Outside Room Reference Phrase	ORMRP
Object Target Phrase	OBTP
Object Reference Phrase	OBRP
Furniture Target Phrase	FURTP
Furniture Reference Phrase	FURRP
Inside Room Reference Phrase	IRMRP
Perspective Indication	PERS
Confusion Indication	CONF

Table 1. List of semantic chunk labels and their abbreviations.

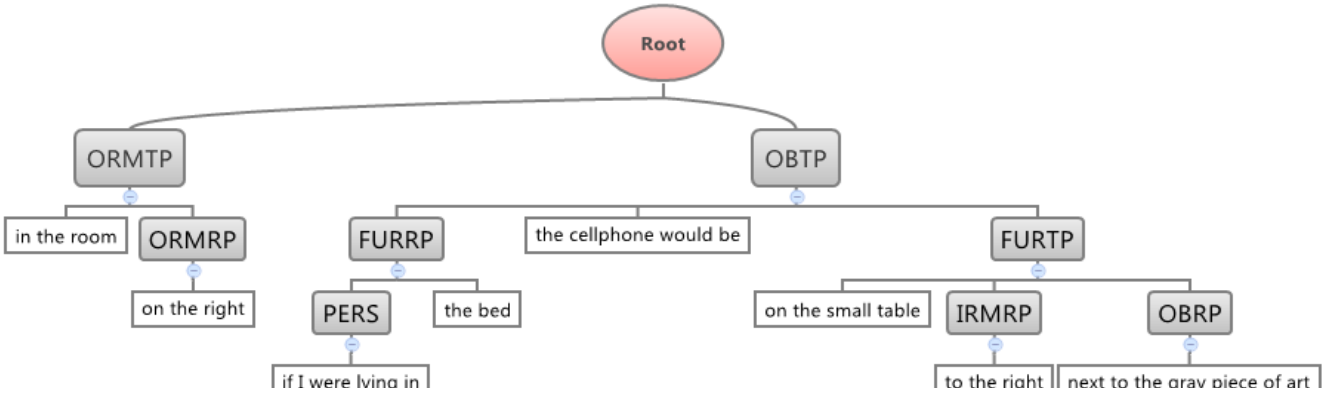

Figure 4. The proposed semantic spatial language grammar applied to a real description (CSISL corpus) from human subject experiment conducted by Carlson and Skubic (2011).

The former is referred to as a "target" component and the latter a "reference" component. These were referred to as "goal" and "path" in the template examples in Figure 3. As mentioned in the semantic PoS section, an assistive robot may need to use different perception methods for small objects, furniture and house structures which motivated the separation between these groups.

The top-most non-root nodes are both "target" components and therefore their chunk labels end with "TP" which stands for target phrase. The "RP" ending of a chunk label stands for "Reference Phrase". There were no Inside Room Target Phrases because descriptions either contained the more specific target furniture landmark or were phrased in a way that did not imply any target such as "turn right". There were some components that did not fit any of the above categorical separations. These consisted of specific "perspective" suggestions such as "as if you're looking at the bed", which were labeled with the "PERS" chunk label and corrections of what was previously said or just indicators of confusion, which were labeled "CONF".

The grammar has rules for deciding which labels can be parents to other labels. Target phrases are always parents of reference phrases. Generally OBTPs will be parents to other phrases (in some cases the rest of the description, including ORMTP) because this is the "goal" of the description. If the description started with "the cellphone is in the room on the right", as many static "where" descriptions did, OBTP would be the parent of every other node. PERS and CONF have to be children of the other chunk types, because "perspective" and "confusion" are always about some other part of the description. In the example in Figure 4, the perspective phrase "if I were lying in" refers to the FURRP "the bed" and is its child. If, for example, the description started with "in the room on the right wait no actually on the left", "on the right wait no actually" would be a CONF child of the ORMRP component "on the left". Another important rule is that the

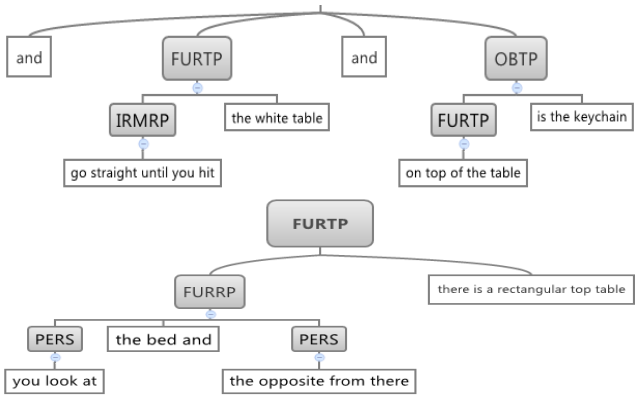

Figure 5. Parts of real descriptions from the corpus that demonstrate when "and" is (bottom) or is not (top) part of a chunk.

prepositions or hedges, i.e. "on, in, next to, to", etc. between two different components are always in the child node. Adverbs such as "then" and conjuncts such as "and" are almost always outside of any chunk or outside of the child chunk. Figure 5 shows examples of each case. The bottom example could not be separated at "and" because then the link between "bed" "opposite from" and "table" would have been lost if they were in separate chunks.

Corpus Annotation

The corpus of 1024 real descriptions as well as the corpus of 149 template descriptions was annotated with PoS tags (including the special semantic tags) and the spatial semantic grammar discussed in this paper using an XML output format. A hybrid approach of manual and automated annotation was used. NTLK (Bird 2004, 2009) and standard Python libraries were used for this task. Any punctuation, if present, was stripped because there is no punctuation when speaking, although it could be added back in with automatic comma and period placement methods.

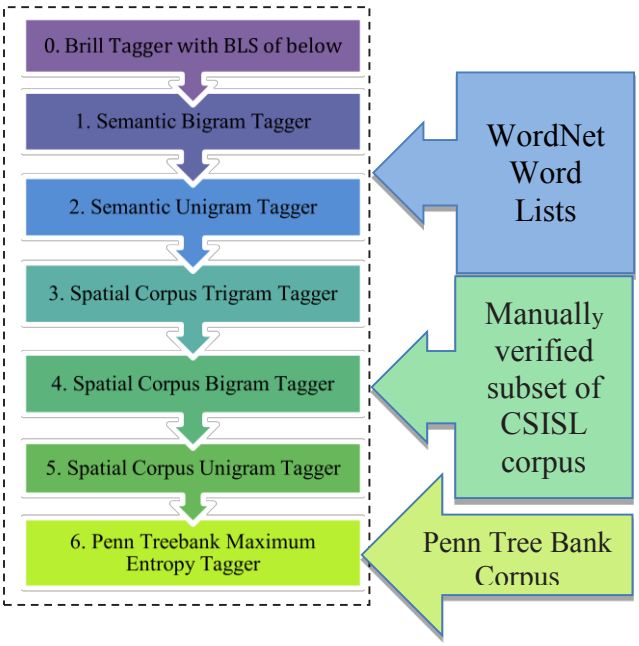

Figure 6. Architecture of the back-off semantic PoS tagger and the source of training data. The Brill tagger uses all the other taggers as a Base Line System (BLS) and therefore does not have its own training data.

Part-of-Speech Tag Annotation

A specialized back-off tagger was created as shown in Figure 6. First the Maximum Entropy Tagger included with NLTK and trained on the Penn Tree Bank Corpus was applied to the untagged spatial language corpus. Then a

subset of the CSISL corpus was manually checked for systematic errors and corrected, which was necessary because the Penn corpus is from a different domain. A unigram, bigram and trigram tagger were trained on the corrected subset and applied to the rest of the corpus.

Next, lists of words for furniture items (FUR) and house structures (STR) and room names (RM) were taken from WordNet (Fellbaum 1998), and unigram and bigram taggers were trained using only these words. When applied to the result of the previous steps, the Semantic Taggers would only change the FUR and STR labels. Finally a Brill Tagger (Brill 1995) was trained using the back-off taggers (1-6 in Figure 6) as a Base Line System. Brill Taggers learn a set of error-correcting rules from templates that specify the context windows for nearby words and tags. This addition usually increases the accuracy by up to 5% (Brill 1995). The resulting PoS tagger was then applied to the rest of the spatial language corpus to provide PoS tag annotation.

Semantic Grammar Annotation

For the semantic grammar annotation, a manual approach was required; however, some automatic techniques were used. To simplify this task and provide a widely-supported format, the PoS-tagged lists (descriptions) were then parsed with a coarse Regular Expression Parser (Bird 2009). The segmentation and labels were mostly incorrect; however, they helped by breaking up the descriptions. The resulting list of Trees was then serialized to an XML format. The incorrect chunk (or Tree) labels were then fixed manually in the XML file. The author first hand-annotated the 149 descriptions in the template corpus and applied the semantic grammar to a few real descriptions to make sure it can be applied to the real descriptions at all. The task of annotating the rest of the descriptions was partially outsourced to two undergraduate Linguistics students. The corpus of 1024 real descriptions (CSISL) was then broken up into 3 parts: 512 for the first student, the other 512 for the second student and an overlap of these two sets consisting of 256 descriptions was annotated by the author for cross validation. The descriptions were sampled randomly into the subsets to ensure that descriptions from different experimental subgroups appeared in each set. The semantic spatial grammar was refined and changed throughout this process, requiring the annotators to go back and re-annotate. While grueling, this led to a grammar that captures the vast majority, if not all, spatial relationships found in the corpus. Figure 7 (left) shows what the end result looked like in XML format after annotation. The annotators were presented with similar input with the exception of incorrect and missing chunk labels (from RegEx Parser) which they needed to fix. The XML tags around PoS labels were provided to the annotators to speed up the process.

```
<Note>
<OBTP>
<DT>the</DT>
<NNS>keys</NNS>
<VBP>are</VBP>
<FURTP>
<ON>on</ON>
<DT>the</DT>
<JJ>white</JJ>
<FUR>stand</FUR>
<ORMTP>
<IN>in</IN>
<DT>the</DT>
<JJ>living</JJ>
<RM>room</RM>
</ORMTP>
</FURTP>
</OBTP>
</Note>
```

"the"	OBTP
"keys"	OBTP
"are"	OBTP
"on"	OBTP-FURTP
"the"	OBTP-FURTP
"white"	OBTP-FURTP
"stand"	OBTP-FURTP
"in"	OBTP-FURTP-ORMTP
"the"	OBTP-FURTP-ORMTP
"living"	OBTP-FURTP-ORMTP
"room"	OBTP-FURTP-ORMTP

Figure 7. (left) A portion of the annotated corpus stored in XML format. (right) Nested Chunk Encoding applied to this example.

Nested Chunk Encoding

The semantic grammar proposed in this paper is nested unlike the traditional "chunks"; the nesting needs to be preserved since it captures useful spatial relationships. This made the commonly used Inside-Outside-Begin (IOB) Encoding (Bird 2009) not usable. At the same time the authors wanted to use methods usually used for chunking or PoS tagging for this HRI application because deeper parsing methods can be slow at runtime and traditional grammar-based parsers require specifying every possible word, tag and chunk and their exact combinations rather than rely on more general contextual and morphological features, which allow extensibility to a different map, for example. This led to the development of the Nested Chunk Encoding which can serialize the nested chunk labels into a single label and preserve the structure. The method is very simple and the serial chunk labels are shown in Figure 7 (right). As shown in Figure 7, the method simply appends the parent chunks to the left of the child chunks with "-"as a separator. A feature can be based on the constituents of the nested chunk in addition to the whole chunk by splitting at the "-". Although it may seem that this is somehow inefficient for deeper nesting that is not the case. The deepest nesting observed in the corpus was 6 labels appended together. There is a natural limit on the depth of nesting since deeper nesting would make the descriptions difficult to follow for humans. It is unlikely anyone would

use deeper levels of nesting, especially in spoken language, unless the purpose was to confuse the addressee. Additionally, there was no noticeable slowdown in the training or testing of the Brill tagger when this encoding was used instead of the IOB encoding.

Results

Per-Word-Accuracy (PWA) results of the automated PoS and Chunking methods are presented in Tables 2 and 3. The results for the PoS tagging are in line with the state-of-the art. Thechunking results cannot be directly compared with any existing results because a novel corpus was used; future work will involve measuring the accuracy of other methods using the CSISL and template corpora.

PoS Tagging Results

Table 2 shows the accuracy results (in %) of different training and testing data combinations. The first row used only the Maximum Entropy Tagger supplied with NLTK which was trained on the Penn Treebank and tested on 624 descriptions from the young adults and 312 from the old. It was observed by Carlson et. al. that the vocabulary used by younger and older adults was different, however PoS tagging accuracy was not affected according to the results.

Training	Testing	PWA	PWA With Brill
Penn Treebank	624 Y + 312 Old	**0.67**	n/a
1024 Young	512 Old	0.96	**0.97**
512 Old	1024 Young	0.95	0.96
400 Y+200 Old	624 Y +312 Old	0.96	**0.97**

Table 2. PoS Tagging accuracy results in % for different combinations of training and testing data.

Nested Chunking Results

Table 3 shows the results for automatic chunking. The Brill tagger (Brill 1995) was used for training an automatic text chunker (Bird 2009) with portions of hand-annotated corpora used as training data. In each case, 80% of the corpus was used for training. The PWA of the chunker (rather than F-score) was evaluated. This metric was more appropriate considering the very high accuracy results for the Template corpus and much lower results for the Real corpus. The F-scores, more commonly reported metrics, would be a lot less informative as they would be near zero for the CSISL corpus due to the large accuracy error rate.

Corpus	Min PWA	Max PWA
Templates	93%	**99%**
CSISL	**54%**	62%

Table 3. Brill Chunking accuracy Min and Max out of 20 trials. 80/20 training testing split.

Discussion

The PoS tagging results were in the upper ninety percentile, which is comparable with state of the art and sufficient for the task. The text chunking results on the template corpus were very high, nearly 100% in some trials and no less than 93%, which is very high considering the nesting, which greatly increases the number of possible labels, and the small corpus size (149 template descriptions). The reason for the relatively low accuracy of the chunking on the real descriptions (CSISL) is a combination of factors. First there are lingering errors and possible (not measured, but noticeable) lack of cross-annotator agreement. The accuracy results of the subset annotated by the author were about 7% higher. Since the annotation was done over the course of 2 months, the annotators became better at the task over time. The Brill tagger may also not be the best method for the task and was chosen for its speed and availability. More complex models combined with voting schemes are likely to outperform Brill at this task just as they do at other NLP-related classification tasks (Collobert 2011). Another problem was that the Brill tagger only operates on doubles, which meant that only (PoS tag, chunk tag) or (word, chunk tag) could be used for training, not the entire triple, leading to loss of potentially useful features. A more detailed analysis of the source of errors (likely rare labels such as “PERS” and “CONF”) needs to be conducted.

An interesting application of the semantic grammar and the template description corpus could be the use of Probabilistic Context Free Grammars (PCFG) to generate complex spatial descriptions from scenes or sets of observations. PCFGs are generative and can both parse a description or use the grammar to generate a description. This method could also be used to create additional training data for the chunking task.

Conclusion

This paper proposes a semantic indoor spatial language grammar rooted in interdisciplinary research capable of capturing complex spatial relationships. A novel method that allows single-label representation of deeply nested trees is also presented, motivated by the needs of an HRI application. Preliminary results are included along with a discussion of possible improvements and additional applications. Although the PWA was low on CSISL corpus, the result is very promising considering the annotation inconsistencies, small training set and the use of a simple chunking method. Future work will aim to fix these problems.

References

Bird, S., and Loper, E. 2004. NLTK: The natural language toolkit. In Proc., *42nd Annual Meeting of the Association for Computational Linguistics* (ACL-04).

Bird, S., Klein, E., Loper, E. 2009. *Natural Language Processing with Python (1st ed.)*. O'Reilly Media, Inc.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4), 543-565.

Carlson, L., Skubic M., Miller J., Huo Z., Li, X. 2011. A Corpus of Spatial Descriptions for the Development of Human-Driven Spatial Language Algorithms. *52nd Meeting of the Psychonomics Soc*.

Carlson L., Skubic M., Miller J., Huo Z. & Alexenko T. 2013. "Assessing the Effectiveness of Older Adults' Spatial Descriptions in a Fetch Task," *Proceedings, 35th Annual Cognitive Science Conference*, Berlin, Germany, pp 281-286.

Clark, H.H., Schreuder, R., and Buttrick, S. 1983. "Common Ground and the Understanding of Demonstrative Reference," *Journal of Verbal Learning and Verbal Behavior*, 22:1-39.

Clark, H.H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12: 2493-2537.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Levit, M. and Roy, D. 2007. "Interpretation of Spatial Language in a Map Navigation Task," *Proc., IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(3):667-679.

Marcus, M., Santorini, B., Marcinkiewicz, M. 1993. "Building a large annotated corpus of English: the Penn Treebank." *Computational Linguistics*, 19(2).

Matuszek, C., Fox, D., and Koscher, K. (2010). "Following Directions Using Statistical Machine Translation," *Proc., Intl. Conf. on Human-Robot Interaction*, Osaka, Japan.

Skubic, M., Alexenko, T., Huo, Z., Carlson, L., Miller, J. 2012. Investigating spatial language for robot fetch commands, *AAAI Technical Report, WS-12-07*, 39-45.

Skubic. M., Carlson, L., Li, X., Miller, J., Huo, Z. 2012. "Spatial language experiments for a robot fetch task." *in Proc., ACM/IEEE Intl. Conf. on Human-Robot Interaction*, Boston, MA.

Skubic M, Huo Z, Alexenko T, Carlson L & Miller J. 2013. "Testing an Assistive Fetch Robot with Spatial Language from Older and Younger Adults,"*IEEE International Symposium On Robot and Human Interactive Communication"*, Gyeongju, Korea.

Skubic. M., Huo, Z., Carlson, L., Li, Z., Miller, J. 2011. "Human-Driven Spatial Language for Human-Robot Interaction." *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Skubic, M., Matsakis, P., Chronis, G. and Keller, J. (2003). "Generating Multilevel Linguistic Spatial Descriptions from Range Sensor Readings using the Histogram of Forces," *Autonomous Robots*, 14(1): 51-69.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S. and Roy, N. 2011. "Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation," *Proc., Conf. on Artificial Intelligence* (AAAI).