Learning to Distinguish PP Arguments from Adjuncts

Aline Villavicencio
Computer Laboratory, University of Cambridge
J.J Thomson Avenue, Cambridge, CB3 OFD, UK
Phone: +44-1223-763642
Fax: +44-1223-334678
Aline.Villavicencio@cl.cam.ac.uk

Abstract

Words differ in the subcategorisation frames in which they occur, and there is a strong correlation between the semantic arguments of a given word and its subcategorisation frame, so that all its arguments should be included in its subcategorisation frame. One problem is posed by the ambiguity between locative prepositional phrases as arguments of a verb or adjuncts. As the semantics for the verb is the same in both cases, it is difficult to differentiate them, and to learn the appropriate subcategorisation frame. We propose an approach that uses semantically motivated preposition selection and frequency information to determine if a locative PP is an argument or an adjunct. In order to test this approach, we perform an experiment using a computational learning system that receives as input utterances annotated with logical forms. The results obtained indicate that the learner successfully distinguishes between arguments (obligatory and optional) and adjuncts.

1 Introduction

Words differ in the subcategorisation frames that realise their semantic arguments, and a given word may have several different subcategorisation frames. The subcategorisation frame includes all the complements of a given word. For instance, the sentences:

- (1) *John ate*
- (2) *John ate the apple*

represent the intransitive and transitive frames, respectively, and both are valid frames associated with the word *eat*. Given that the subcategorisation frame of a given word should only include a given constituent if it is an argument, one problem is caused by the ambiguous nature of some constituents, that can be either arguments or adjuncts.

The ability to distinguish between subcategorised arguments and non-subcategorised adjuncts is of great importance for several applications, such as automatic acquisition of subcategorisation lexicons from data, and this problem has been widely investigated. For instance, Buchholz (1998) investigates this task using a memory-based learning approach, where the use of syntactic and contextual features results in a 91.6% accuracy in distinguishing arguments from adjuncts. Brent (1994) looks at the problem from a more psychologically oriented perspective, trying to simulate the environment available to a human language learner, and using binomial error estimation to derive subcategorisation frames for verbs, based on imperfectly reliable local syntactic cues. This technique is able to capture the fact that the relative frequency of a verb-argument sequence is likely to be higher than that of a verb-adjunct sequence. However, the cues used in the simulations are too simple to achieve high accuracy. Steedman (1994) suggests the use of semantic information to deal with this ambiguity, given that syntax should be as close as possible to semantics. Then, given that for a particular language there is a strong correlation between the subcategorisation frames and predicate-argument structure of a given word, from the predicate-argument structure of a word it is possible to infer its subcategorisation frame.

In terms of the difficulty of this task, Buchholz (1998) found that in the experiments conducted the ambiguity presented by Prepositional Phrases (PPs) was the most difficult case to classify, accounting for 23% of the errors. Moreover, Brent (1994) also found in his simulations that locative adjuncts were sometimes mistaken for arguments. In this paper we focus on the problem of distinguishing between locative PPs as arguments or adjuncts, where only if a given locative PP is an argument is that it should be included in the subcategorisation frame of the verb. The approach proposed here is to use semantically motivated preposition selection and frequency information to determine if a locative PP is an argument of the verb or if it is an adjunct. In order to test this approach, we use a computational learning system, and the results obtained indicate the effectiveness of the approach.

The wider goal of this project is to investigate the process of grammatical acquisition from data. Thus, in section 2 we start by giving some background in language acquisition employed in the learning model, which is described in section 3. Characteristics of the ambiguity between arguments and adjuncts are discussed in section 4 together with the approach used to distinguish them. In section 5 we describe an experiment conducted to test the approach. We finish with some conclusions and a discussion of future work.

2 Language Acquisition

In trying to solve the question of how to get a machine to automatically learn language from data, we can look at the way people do it. When we acquire our mother language we are exposed to an environment that includes noisy and ungrammatical sentences, the potential influence of other languages, and many other linguistic phenomena. In spite of that, most children are successful in the acquisition of a grammar in a relatively short time, acquiring a sophisticated mechanism for expressing their ideas, based on data that is said to be too impoverished to generate such a complex capacity. One approach to explain the acquisition of languages proposes that children must have some innate knowledge about language, a Universal Grammar (UG), to help them overcome the problem of the poverty of the stimulus and acquire a grammar on the basis of positive evidence only (Chomsky 1965). According to Chomsky's Principles and Parameters Theory (Chomsky 1981), the UG is composed of principles and parameters, and the process of learning a language is regarded as the setting of values of a number of parameters, given exposure to this particular language. Another likely source of information that is available to children when learning a language is the semantic interpretation or related conceptual representation. Indeed, as Steedman (1994) puts it:

"*Since the main thing that syntax is for is passing concepts around, the belief that syntactic structure keeps as close as possible to semantics, and that in both evolutionary and child language acquisition terms, the early development of syntax amounts to little more than hanging words onto the preexisting armatures of conceptual structure is so simple and probable as to amount to the null hypothesis*".

A third source of information can be found in the statistical properties of the input data to which children seem to be sensitive, as observed in recent work in psycholinguistics.

3 The Learning System

These ideas about human language acquisition are employed, in this work, in the construction of a computational learning system that can learn from its linguistic environment, which may contain noise and ambiguities (Villavicencio 2002).

Studies like this can not only be used to provide clues about possible directions to follow in the automatic acquisition of information from data, but also to help us understand better the process of human language learning. However, if that is to be achieved, we need to concentrate only on algorithms and resources that a human learner could employ. Thus, there are significant constraints on the assumptions that can be made in the learning system implemented. In this way, the learner cannot have access to negative information; it also cannot start with information specific to a particular language, and can only assume information that is general among human languages. Another aspect is that learning has to be on-line and incremental, with the system only processing one sentence at a time, without the possibility of storing sentences and reprocessing previously seen sentences, or doing multiple passes through the corpus. Moreover, the kind of data given to the learner must be compatible with the linguistic environment of a child.

In this work the linguistic environment of the learner is simulated to a certain extent by using spontaneous child-directed sentences in English, which were extracted from the Sachs corpus (MacWhinney 1995) (Sachs 1983). Some of the semantic and contextual information available to children is introduced in the corpus by annotating the sentences with logical forms. At the moment around 1,500 parents' sentences are annotated with the corresponding logical forms.

The computational learning system employed in this investigation is composed of a UG and associated parameters, and a learning algorithm (Villavicencio 2002). The UG is represented as a Unification-Based Generalised Categorial Grammar, and it provides the core knowledge about grammars that the learner has. A learning algorithm fixes the parameters of the UG to the target language based on exposure to it. In this work, this is in the form of the annotated parents' sentences to simulates some of the characteristics of the environment in which a child acquires her language. Finally, children's sensitivity to statistical properties of the data is also simulated to some extent in the learning system.

4 Learning from Ambiguous Triggers

The learning environment to which the learner is exposed contains noise and ambiguity and the learner has to be able to deal with these problems if it is to set its parameters correctly and converge to the target grammar. In this work we concentrate on the ambiguity in the form of locative PP that can occur either as arguments to a verb or as adjuncts.

When processing a sentence the learner needs to determine appropriate syntactic categories for the semantic predicates used as input in order to correctly set its parameters. In most cases, the learner is able to find the required syntactic categories, using the Categorial Principles (Steedman 2000). According to these principles from the semantic interpretation of a word and some directional information for a language, it is possible to determine the syntactic form of the corresponding category. [1] These principles help the learner to determine the subcategorisation frame for a given word based on its semantic predicate. Then, for instance, in the sentence:

•(3) *John talks to Mary*

with logical form

•(4) talk-communicative-act(e,x,y), john(x), comm-to(y), mary(y)

the verb *talks* has two arguments, the NP subject *John*, and the PP *to Mary*, as represented in the logical form associated with the verb, where the PP is the second argument and as such should be included in the subcategorisation frame of the verb: (S\NP)/PP. On the other hand, in the sentence:

•(5) *Bob eats with a fork*

with logical form

•(6) eat-ingest-act(e,x), bob(x), instr-with(e,y), a(y), fork(y)

the PP *with a fork* is not an argument of the verb *eat* as reflected in its logical form and should not be included in its subcategorisation frame, which is S\NP.

It means that from the logical form associated with a verb, the learner can decide whether a given constituent is an argument of the verb, and should be included as its complement in the subcategorisation frame or not. However, one exception to this case is that of verbs occurring with locative PPs, which can be either arguments or adjuncts. The ambiguity between these cases arises because in this logical form representation the logical form describing the verb with an argument locative PP is similar to that describing the verb with an adjunct locative PP. For example, the sentence:

• (7) *Bill kisses Mary in the park*,

with logical form:

•(8) kiss-contact-act(e,x,y), bill(x), mary(y), loc-in(e,z), the(z), park(z)

exemplifies a case where the locative PP is an adjunct. Thus it should not be included in the subcategorisation frame of the transitive verb *kiss*, which is (S\NP)/NP. On the other hand, the sentence:

•(9) *Bill swims across the river*

with logical form:

•(10) swim-motion-act(e,x), bill(x), motion-across(e,y), the(y), river(y)

shows a case where the PP is an (optional) argument of the verb *swim*, and where the appropriate subcategorisation frame for the verb should include it ((S\NP)/PP), even though the PP is

[1] These principles are closely related to the Projection Principle (Chomsky 1981) that states that the selectional requirements of a word are projected onto every level of syntactic representation.

not included in the logical form of the verb.

For both sentences, the logical form has a similar structure, with both a verbal and a locative predicate, with the PP not being included in the logical form of the verb. As a consequence, the logical form cannot be used to help the learner resolve the ambiguity: given the logical forms {kiss-contact-act(e,x,y), loc-in(e,z)} and {swim-motion-act(e,x), motion-across(e,y)}, which syntactic category should the learner choose for each of these verbs? This ambiguity constitutes a significant problem for the learner, since it has to decide whether a given PP is functioning as a complement of a verb or if it is working as an adjunct. Three different cases to which the learner is exposed are identified, based on Pustejovsky (1995) and Wechsler (1995), with the PP occurring as an obligatory argument, as an optional argument, or as an adjunct[2]:

1. **The PP is an obligatory argument of the verb.** For certain verbs the PP is an obligatory argument of the verb and should be included in its subcategorisation frame. An instance of this case is the verb *put*, in sentence 11:

   - (11) *Mary put the book on the shelf*,

   where the verb occurs with a locative PP. Also, as the ungrammaticality of sentence 12 suggests, this verb requires a locative PP:

   - (12)* *Mary put the book*

   The appropriate syntactic category for the verb[3] is ((S\NP)/PP)/NP.

2. **The PP is an optional semantic argument of the verb.** For example, a verb such as *swim* can occur as in sentence 9, where it is modified by a directional PP which is an optional argument of the verb, but this verb may also occur without the PP, as in sentence 13:

   - (13) *Bill swims.*

   This is a case of a verb that can occur in both constructions with the PP being a semantic argument, which, when occurring, must be included in the subcategorisation frame of the verb. Consequently, the appropriate category for the verb *swim* in sentence 9 is (S\NP)/PP, and in 13 is S\NP.

3. **The PP is an adjunct.** Adjuncts modify the logical form of the sentence, but are not part of the subcategorisation frame of the verb. The PP *in the park* in sentence 7 is an example of an adjunct that is neither part of the semantic argument structure of the verb *kiss* nor part of its subcategorisation frame. This verb can also occur without the PP, as in sentence 14:

   - (14) *Bill kisses Mary.*

   The appropriate syntactic category for the verb in both sentences is (S\NP)/NP.

When faced with a locative PP, the learner has to identify which of these cases is appropriate. The required subcategorisation frame is determined independently for each verb sense, depending on the semantic type of the verb, and on its frequency of occurrence with a particular subcategorisation frame and predicate argument-structure combination.

In order to determine if a locative PP is an obligatory argument of the verb, the learner uses frequency information about the occurrence of each verb with locative PPs. If the frequency with which they occur together is above a certain threshold, the PP is considered to be an obligatory argument of the verb and included in its subcategorisation frame. In this case, the threshold is set to 80% of the total occurrences of a verb. This is high enough for discarding adjuncts and optional arguments that occur occasionally, and at the same time is not high enough to be affected by the occurrence of noise.

In an analysis of all the mother's sentences in the entire Sachs corpus, only two occurrences of *put* without the locative PP were found: one seems to be an instance of an elliptical construction, and the other a derived sense. The frequency with which *put* occurs with a locative PP correctly indicates that the PP is an argument of the verb, and it needs to be included in the subcategorisation frame of the verb. On the other hand, for verbs like *kiss* and *swim* in

[2]In this work we classify PPs in terms of these three cases, even though more fine-grained classifications can be used as by Pustejovsky (1995).

[3]This work does not include, in its investigation, elliptical or noisy constructions. Therefore, the sentences analysed and the frequencies reported exclude these cases.

sentences like 7 and 9, the locative PP is an occasional constituent, with the semantics of the sentence including the location predicate only in these cases. The occasional occurrence of PPs with these verbs correctly indicates that they are not obligatory arguments of the verbs.

If the frequency of occurrence is not above the threshold, then the PP can be either an optional argument or an adjunct. To determine if a PP is an optional argument, the learner uses information about the kind of semantic event denoted by the verb. As Steedman (1994) notes

"... *if we are asking ourselves why children do not classify* ***meet*** *as subcategorising for NP PP on the basis of sentences like (1b),* ***we met Harry on the bus****, then we are simply asking the wrong question. A child who learns this instance of the verb from this sentence must start from the knowledge that the denoted event is a meeting, and that this involves a transitive event concept*".

Thus, when the learner receives an input sentence, it uses semantic information about the kind of event denoted by the verb and preposition given in the logical form associated with the sentence to check if the preposition can be selected by the verb. This approach to identify non-obligatory argument PPs is based on Wechsler's proposal of semantically motivated preposition selection (Wechsler 1995), where a PP is an argument of a verb if it can be selected by the verb on pragmatic grounds. The learner represents pragmatic knowledge in terms of a hierarchy of types and words are classified according to these types, based on the semantics associated with them.[4] A verb can select a preposition as an argument if the latter is of the same type as the verb, or of one of its subtypes in the hierarchy. A fragment of such a hierarchy is shown in figure 1. Then, a verb such as *talk* (in *John talks to Mary*), which as specified in the logical form (in 4) denotes a communicative event and is an instance of type **communicative-act**, can select as its optional argument a preposition such as *to*, which is of type **comm-to**, because the latter type is a subtype of the former on the world knowledge

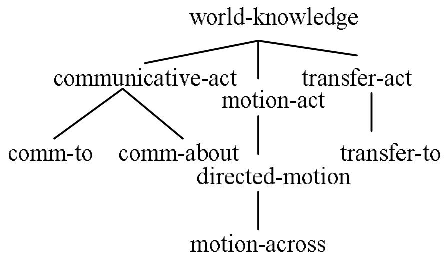

Figure 1: Fragment of Hierarchy of World Knowledge

hierarchy. On the other hand, this verb does not select a preposition such as *across* of type **motion-across** as its argument, in a sentence such as *Bill talked about his treatment across the country*, because their types do not unify. In this case, the PP is an adjunct to the verb. However, this preposition can be selected as the argument of the verb *swim* in sentence 9, which denotes a motion event and is an instance of type **motion-act**. As words are associated with types in the hierarchy, the lower in the hierarchy a given word is, the more constrained its selectional possibilities are (as discussed by Wechsler (1995)). In this way, the pragmatic knowledge confirms certain PPs as arguments of some verbs, while rejecting others.

If a locative PP is rejected as argument of a verb on pragmatic grounds, then the PP is treated as an adjunct and is not included in the subcategorisation frame of the verb. Once the learner decides which is the case for a particular verb PP combination, it uses the triggering information, including the appropriate subcategorisation frame of the verb, for further processing.

5 Argument or Adjunct?

To test this approach we conducted an experiment where the learner is evaluated in terms of three different verbs: *put* where the PP is an obligatory argument, *come*, where the locative PP is an optional argument, and *draw* (in the sense of drawing a picture) where the PP is an adjunct. These verbs are representative of each case and the sentences in which they occur are taken from the mother's section of the complete Sachs corpus, which is the largest of the parents' sections. The status of the locative PPs

[4] In this work we do not address the issue of how such a pragmatic hierarchy would be constructed and we assume that it is already in place. However, for a related task, see Green (1997).

occurring with these verbs is determined following syntactic tests for argument structure. The specific test used in this case is the "do so" test, which is a standard test for argument structure, as discussed by Verspoor (1997). In this test, the claim is that a complete complement can be replaced by "*do so*". In the case of obligatory arguments, only the full constituent **verb PP** or **verb NP PP** can be replaced by *do so*, while in the case of adjuncts, the **verb** or **verb NP** constituent can also be replaced by *do so*. Sentences 15 to 23 indicate that the PPs are arguments of the verbs *put* and *come*, and adjuncts of the verb *draw*.

- (15) *You put Goldie through the chimney*
- (16) *You put Goldie through the chimney and Bob also did so*
- (17) ** You put Goldie through the chimney and Bob did so through the window*
- (18) *You came from the garden*
- (19) *You came from the garden and John also did so*
- (20) ** You came from the garden and John did so from the kitchen*
- (21) *You drew in the park*
- (22) *You drew in the park and John also did so*
- (23) *You drew in the park and John did so in the garden*

In these experiments, the learner is given as input sentences from the annotated Sachs corpus (all previously unseen), among which are the ambiguous cases, as shown in table 1, collected from the mother's corpus. The learner processes each sentence, having to determine valid syntactic category assignments for each word in the sentence (Villavicencio 2002) (Waldron 2000), and based on these, setting the parameters of the UG. For each sentence the learner collects information about the words, their corresponding logical forms, syntactic categories, and frequency of occurrence. When the learner is faced with an ambiguous sentence, it needs to disambiguate the PP as argument or adjunct of the verb. It first checks if the frequency of occurrence of the verb with locative PPs as seen so far is above the threshold of 80%, in which case the PP is considered to be an obligatory argument of the verb. Otherwise, the learner checks if the verb can select the PP on pragmatic grounds, based on the pragmatic hierarchy the learner has, and on the logical form associated with the words. If so, the PP is an optional complement of the verb. On the other hand, if this is not the case, then the PP is considered to be an adjunct. After deciding, the learner proceeds with the setting of parameters and collects the new frequencies, as described above, and goes on to process the next sentence.

Table 1: Disambiguation of Locative PPs

Verb	Sentences with PPs	Total Sentences
put	137	137
come	24	32
draw	9	21

The results obtained for each of these three verbs are that the learner correctly selects the appropriate subcategorisation frame in all of these cases, which confirms the effectiveness of the proposed approach to disambiguate locative PPs. In terms of frequency of occurrence of the verbs with the locative PPs, other verbs in the mother's sentences from the entire Sachs corpus also have a similar pattern, with the locative PP being frequent for obligatory arguments of the verb, and less frequent for the other cases:

- *stay*, which according to the "do-so" test has an obligatory locative PP argument, occurs in 100% of the cases with locative PPs,
- *come*, which has optional locative PP arguments, occurs in 69.6% of the cases with locative PPs, and all of these can be semantically selected by the verb,
- *eat*, as a transitive verb which does not have a locative PP argument, occurs in only 1.23% of the cases with locative PPs, and
- *play*, as an intransitive verb also does not have a PP argument, is in 40% of the cases with a locative PP.

These results indicate that the proposed approach indeed helps the learner to disambiguate between locative PPs as arguments or adjunct based on frequency information and semantically motivated selection. Such an approach provides a possible way forward in which to deal with this problem for the research in the area. It follow Steedman's suggestion about the use of semantic information, and similarly to Brent and Buchholz it uses local information to deal with this ambiguity, in a setting that is compatible with some studies on language acquisition.

6 Conclusions and Future Work

In this paper we described one possible approach to deal with the problem of disambiguating between arguments or adjuncts. This approach is tested by a learning system used to investigate the automatic acquisition of language from data. The learning system is equipped with a plausible model of the Universal Grammar and it has to set its parameters to the target language based on the input data. The ambiguity between arguments and adjuncts is one of several difficulties encountered by the learning system during the acquisition process and the approach proposed to overcome this problem, proved to be effective and helped the learner decide the appropriate case for the ambiguities found in the data available. The implemented learning system can successfully learn from a corpus of real child-directed data, containing noise and ambiguity, in a more realistic account of parameter setting (Villavicencio 2002).

Disambiguation based on frequency information and semantically motivated selection provides a plausible strategy, compatible with some research on language acquisition. Although this is primarily a cognitive computational model, it is potentially relevant to the development of more adaptive NLP technology, by indicating possible paths for future developments in the area. However, larger scale tests still need to be conducted to see how the approach would generalise, and for that we need more annotated data. These two tasks of annotating more data and undertaking this larger scale investigation are included in the future directions of this work.

Acknowledgments

Thanks to Francis Bond for comments on this paper. This research was supported in part by the NTT/Stanford Research Collaboration, research project on multiword expressions, and by CAPES/Brazil.

References

Michael R. Brent. *Surface Cues and Robust Inference as a Basis for the Early Acquisition of Subcategorization Frames.* In L. Gleitman and B. Landau eds. The Acquisition of the Lexicon, 1994.

Sabine Buchholz. *Distinguishing Complements from Adjuncts Using Memory-Based Learning.* In B. Keller ed. 'Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing, 1998.

Noam Chomsky. *Aspects of the Theory of Syntax.* MIT Press, 1965.

Noam Chomsky. *Lectures on Government and Binding.* Foris Publications, 1981.

Georgia M. Green. *Modelling Grammar Growth: Universal Grammar without Innate Principles or Parameters.* Unpublished manuscript prepared for GALA97 Conference on Language Acquisition, Edinburgh, 1997.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk.* Second Edition, 1995.

James Pustejovsky. *The Generative Lexicon.* MIT Press, 1995.

Jacqueline Sachs. *Talking about the There and Then: the Emergence of Displaced Reference in Parent-Child Discourse.* In K. E. Nelson editor, Children's language, v.4, 1983.

Mark Steedman. *Acquisition of Verb Categories.* In L. Gleitman and B. Landau eds. The Acquisition of the Lexicon, 1994.

Mark Steedman. *The Syntactic Process.* The MIT Press, 2000.

Cornelia M. Verspoor. *Contextually-Dependent Lexical Semantics.* PhD Thesis, University of Edinburgh, 1997.

Aline Villavicencio. *The Acquisition of a Unification-Based Generalised Categorial Grammar.* PhD Thesis, University of Cambridge, 2001. Available as Technical Report N. UCAM-CL-TR-533, 2002.

Ben Waldron. *Learning Natural Language within the Framework of Categorial Grammar.* Proceedings of the Third CLUK Colloquium, 2000.

Stephen Wechsler. *The Semantic Basis of Argument Structures.* CSLI Publications, 1995.