Title
Semantic Classifcation of verbs from their Syntactic Contexts: Automated Lexicography with Implications for Child Language Acquisition

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 12(0)

Author
Brent, Michael R.

Publication Date
1990

Peer reviewed

Semantic Classification of Verbs from their Syntactic Contexts: Automated Lexicography with Implications for Child Language Acquisition

Michael R. Brent
MIT AI Lab
545 Technology Square
Cambridge, Massachussets 02139
michael@ai.mit.edu

Abstract

Young children and natural language processing programs share an insufficient knowledge of word meanings. Children catch up by learning, using innate predisposition and observation of language use. However, no one has demonstrated artificial devices that robustly learn lexical semantic classifications from example sentences. This paper describes the ongoing development of such a device. An early version discovers verbs with a non-stative sense by searching in unrestricted text for verbs in syntactic constructions forbidden to statives. Our program parses unrestricted text to the extent necessary for classification. Once the parsing is done recognizing the telltale constructions is so easy even a two-year-old child could do it. In fact, Landau and Gleitman (1985) and especially Gleitman (1989) argue that children must, can, and do use the syntactic constructions in which verbs appear to support meaning acquisition. In this paper we use our program to examine the difficulty of exploiting two particular syntactic constructions to discover the availability of non-stative senses, concluding that only very little sophistication is needed. This conclusion bolsters the general position of Gleitman (1989) that children can exploit syntactic context to aid in semantic classification of verbs.

1 Introduction

Young children and natural language processing programs face a common problem: everyone else knows a lot more about words. Children indisputably catch up by learning, using innate predisposition and observation of language use. However, no one has succeeded in creating artificial devices that robustly learn lexical classifications from example sentences. This paper describes the ongoing development of such a device. An early version discovers verbs with a non-stative sense by searching in edited text[1] for verbs in syntactic constructions forbidden to statives. To do this, it must partially parse the sentence, which in turn requires knowing the major syntactic categories of most of the words. Once the partial parsing is done, recognizing the telltale constructions is so easy even a two-year-old child could do it. In fact, Landau and Gleitman (1985) and especially Gleitman (1989) argue that children must, can, and do use the syntactic constructions in which verbs appear as an aid to meaning acquisition.[2] In the first part of this paper we use our program to examine the difficulty of using two particular syntactic constructions to discover the availability of non-stative senses. We conclude that only very modest syntactic and lexical capability are needed to exploit observation of these syntactic constructions for lexical semantic classification of verbs. This conclusion bolsters the general position of Gleitman (1989) that children can exploit the syntactic structure to aid in semantic classification of verbs.[3]

The second part of this paper describes current work on expanding the scope of the acquisition program both in terms of the lexical classifications it can learn and the quantity and type of text it can learn from. One of several benefits of this extension will be the ability to apply our learning program to corpora of maternal speech, thereby bringing it closer the child acquisition questions.

We have argued for the pursuit of automatic lexical classification based on syntactic context as it relates to questions about children's lexical acquisition, but we would also like to motivate it as technology.

There is wide agreement that language users, whether natural or artificial, need detailed semantic and syntactic classifications of words. However, most current approaches to satisfying that need for artificial devices do not involve learning from examples. Interpreting the information published in machine-readable dictionaries (e.g. Boguraev and Briscoe, 1987), entering it manually in conjunction with knowledge-base construction (Knight, 1989), and studying word collocations statistically (Church and Hanks, 1990) are alternative pro-

[1]The text source we've used so far is the Lancaster/Oslo/Bergen (LOB) Corpus, a balanced corpus of one million words of British English. All but a small percentage is edited prose.

[2]See Pinker (1984) and Pinker (1989) for a different perspective how child learners use constraints between lexical syntax and lexical semantics.

[3]We do not intend to claim that child learners exploit our particular syntactic constructions, which merely represent our earliest attempts; instead, we mean to model the general process of meaning classification based on the syntactic environments in which verbs appear.

posals. When learning from examples is proposed, it is usually tutored learning in a controlled environment (Zernik and Dyer, 1987). Ultimately, however, any language user must be able to add new words to its lexicon, if only to accommodate the many neologisms it will encounter. Moreover, researchers rarely agree on the necessary lexical classifications, and our lexicographic needs grow with our understanding of language. Any method that requires explicit human intervention — be it that of lexicographers, knowledge engineers, or "tutors" — will lag behind both the growth of vocabulary and the growth of linguistics, as well as being subject to the uncertainties of introspection. Further, the cost of maintaining dictionaries manually in the face of this growth will remain high. By contrast, dictionaries constructed by automated learners from real sentences will not lag behind vocabulary growth; examples of current language use are free and nearly infinite. And the ability of such dictionaries to keep pace with theoretical developments is limited only by the difficulty of coming up with syntactic tests and programming the system to detect revealing sentences. Judging by our experience so far, the former task will be the more challenging one.

2 Detecting Verbs with Non-Stative Senses

In this section we discuss our study of two syntactic constructions that reveal the availability of non-stative senses for verbs. This work focuses on three questions to determine the difficulty of discovering the availability of non-stative senses:

1. Is it possible to robustly detect sentences of the type illustrated in (1) and (2) using only a simple syntactic parse tree? How simple can the parse tree be?

2. Do the progressive and rate adverb tests actually behave as advertised in text, which is subject to performance limitations? Specifically, are the syntax-semantics constraints regular enough to support semantic classification under a broad range of psychologically plausible learning strategies?

3. Can a parse tree sufficient for Item 1 be reliably, automatically recovered using only a simple parser, a relatively compact grammar description, and knowledge of the major syntactic categories of the words involved?

Section 2.1 describes the two syntactic constructions we have studied and demonstrates their relation to the semantic category in question. Sections 2.2, 2.3, and 2.4, respectively, answer the three questions in the affirmative. Section 2.5 briefly describes the mechanism and resources used by the parser behind our lexical semantic learning program.

2.1 Revealing Constructions

The distinction between stative and non-stative verbs has been a subject of interest in linguistics at least since Lakoff (1965). Giving a precise semantic characterization of statives is rather involved (see Dowty, 1979); but, roughly speaking, they are verbs that, when asserted at some time, are assumed by default to hold at all later times. Classic examples of stative verbs are *know*, *believe*, *desire*, and *love*. A number of syntactic tests have been proposed to distinguish between statives and non-statives (again see Dowty, 1979). For example, stative verbs cannot normally appear in the progressive, (1). In

(1) a. OK Jon is fixing his car
    b. * Jon is knowing calculus

addition, statives cannot be modified by rate adverbs such as *quickly* and *slowly*, (2). We have chosen the stative/non-stative distinction, and in particular the

(2) a. OK Jon fixes his car quickly
    b. * Jon knows calculus quickly

two tests shown (1) and (2), as test cases for our approach to learning lexical semantic classifications from the syntax of example sentences.

2.2 Required Precision of Parse Trees

Consider first how much syntactic structure is needed to detect the progressive and rate adverb constructions. To begin with, let us assume that the availability of a non-stative sense is an intrinsic property of a verb independent of factors such as the subcategorization frame in which it appears.[4] To detect progressives one need only parse the auxiliary system. Rate adverbs, by contrast, require determining what the adverb modifies. For example, adverbs may appear after the direct object, (3a), and this must not be confused with the case where they appear after the subject of an embedded clause, (3b).

(3) a. Jon fixed the robot quickly
    b. Jon knew his hostess rapidly lost interest in such things

Thus it is necessary to determine the boundaries of NPs and Ss. However, it is not necessary to know much about the internal structure of these phrases.

[4] This is a reasonable approximation for the stative/non-stative distinction. (The obvious exceptions are verbs like *think* that are stative with a propositional argument.) Subcategorization frames are essential for determining many other lexical semantic categorizations. See Section 3.

For example, it is not necessary to know the structure of noun-noun predications, or (except in contrived cases) the attachment of PPs. Finally, there are some truly ambiguous cases that cannot be resolved by any syntactic parser not already possessed of the distinctions we are attempting to learn, (4). When

(4) a. Jon fixed the robot that had spoken slowly
b. Jon believed the robot that had spoken slowly

encountering such sentences the strictly monotonic learner must recognize the ambiguity and decline to draw any conclusion. In summary, to the question: "Is it possible to detect sentences containing constructions (1) and (2) using only a simple syntactic parse?" we answer cautiously, "Yes, in principle."

2.3 Behavior of Test Constructions in the LOB

We now proceed to the question of whether the syntactic tests behave as advertised in real text, which is subject to the hazards of our linguistic performance as well as the rigors our competence. This question is addressed in two stages: first, we estimate the reliability of the progressive and rate-adverb constructions as indicators of non-stativity; next, we evaluate the implications of the reliability estimates for a broad range of psychologically plausible learning strategies. To investigate the reliability of our two constructions as indicators of non-stativity we processed the Lancaster/Oslo/Bergen (LOB) corpus of one million words of edited British English text. After partially parsing each sentence, our program automatically examined the parse trees and the words to see whether or not each clause in fact contained a progressive verb or a verb modified by a rate adverb. When the clause was determined to contain such a construction our program noted that in its dictionary entry for the appropriate verb. It also stored the example sentence for analysis by the researchers.[5]

Before considering the reliability estimates, note that we are exploring syntactic constructions that imply the availability of a **non-stative** sense for a verb, but not constructions that imply the availability of a **stative** sense. Accordingly, in conjecturing that a verb has a non-stative sense, we are quite concerned with false positives. Since we are not currently attempting to conjecture that verbs lack a non-stative sense there is no false negative case.[6]

If the progressive and rate-adverb tests behave ideally then every verb that shows up in one of these constructions ought to have a non-stative sense. To estimate how closely the test constructions approach that ideal, we examined by hand the observations our program recorded for the 100 verbs that occur

[5]Storing sentences is not, of course, required for the learning models under consideration here.

[6]After a thorough statistical analysis of our data we hope to explore the possibility of making negative conjectures stochastically. While such stochastic learning may prove valuable technologically, it is less relevant to the current focus.

most frequently in the LOB corpus. These 100 verbs occur about 50,000 times in the corpus, accounting for 50% of all verb occurrences. Of these 100 verbs, 89 occurred in either progressive or rate-adverb constructions at least once, according to our parser. Of the 89 appearing in the test constructions, only one verb, *mean*, lacks a non-stative sense. Of the 100 verbs, a total of six lack non-stative senses: *know, seem, mean, like, believe,* and *understand*. The lone false positive, *mean*, appears in a non-stative construction only once, in the anomalous sentence "It's a stroke, that was what he was meaning." The six verbs lacking non-stative senses appeared 3,835 times in the corpus with only this single occurrence in the progressive and no occurrences modified by rate adverbs.

Now we must translate the estimates of the reliability of the syntactic tests into conclusions about the reliability of various learning strategies that might employ them. The most obvious learning strategy, and the only one we have implemented so far, is to conclude that a verb has a non-stative sense immediately and irrevocably as soon as one of the telltale constructions is seen, independent of what has been seen in the past. This strategy requires no resources on the part of the learner. In particular, it requires no storage or counting of examples and no inference of any kind. Using this strategy, a learner exposed to our one-million word corpus would have mis-classified one in 89 of our verbs as non-stative, given six opportunities. However, it is obviously unrealistic to apply such a learning strategy over an unbounded body of examples — eventually any verb will show up in an anomalous context. Indeed, it is well-known that children retract overgeneralizations at many points in their language development (Brown, 1973). Nonetheless, the reliability of the syntax/semantics correlation in our sample was so strong that even the null strategy provided nearly one-hundred percent accuracy. Given only modest, psychologically plausible strategies and resources a learner could be expected to achieve one-hundred percent accuracy on the sample data. Although a more articulated and realistic learning strategy is beyond the scope of this paper, we hope to develop one shortly.

2.4 The Sufficiency of the Parser

In Section 2.2 we discussed how much structure must be imposed on sentences if the progressive and rate-adverb constructions are to be detected. In Section 2.3 we determined that the progressive and rate-adverb constructions are indeed reliable indicators of the availability of a non-stative sense. In this section we discuss the accuracy with which we can recover the parses deemed necessary in Section 2.2.

The question at hand is whether a sufficient parse tree can be reliably, automatically recovered using only a simple parser, a relatively compact grammar description, and knowledge of the major syntactic categories of the words involved. As mentioned above, we automatically parsed the LOB corpus. More precisely, we parsed each sentence to the extent necessary to determine whether or not each clause in fact contained a progressive verb or a rate adverb. Let us consider the accuracy of our parser/analyzer in terms of the all-important measure of false positives. In other words, let us consider how many verbs that were judged to be either progressive or modified by a rate adverb in fact were not. It is not practical to check manually every verb occurrence that our program judged to be progressive. Instead, we checked 300 such sentences selected at random from among the most commonly occurring verbs. This check revealed only one sentence that was not truly progressive. That sentence is shown in (5a). Recognizing pseudo-cleft constructions like (5a) would require some

(5) a. *go*: What that means in this case **is going** back to the war years...
   b. *see*: The task was solely **to see how speedily** it could be met...
   c. *compare*: ...the purchasing power of the underdeveloped countries in the commonwealth **will rise slowly compared** with that of Europe.

thought and might entail additions to our grammar description or tree-analyzer that go beyond the merely cosmetic.[7] By contrast with the progressives, rate adverbs are infrequent enough that we were able to verify manually all 281 cases our program found. In four of those cases the rate adverb actually modified a verb other than the one that our program chose. Three of these four cases had the structure of (5b), where a wh- relative is not recognized as signaling the beginning of a new clause. This reflects an oversight in our grammar description that can be corrected trivially. The one remaining case of a mis-attributed rate adverb, (5c), would again require some attention to correct. The rate of false positives in sentence detection, then can be estimated at about one serious hazard in 300 for both tests. However, none of these wrongly identified sentences occurred with verbs lacking a non-stative sense, so none resulted in false positives in word classification.

2.5 Mechanism of the Parser

Having discussed the sufficiency of our parser for the task at hand, it is worth mentioning briefly the nature of the parser. It is a local, heuristic parser developed by DeMarcken (1990). De Marcken's parser is local in that its decisions about the boundaries of phrases are independent of its decisions about their attachment. For example, a grammar writer might choose to build PPs but not attach them. Whether or not they are attached does not affect the legitimacy of phrases or their boundaries. De Marcken's parser is heuristic in that the descriptions of legal structures and the decisions about where to look for those structures are decoupled. In other words, if the only NPs that are

[7] A gross but effective solution would be merely to ignore all sentences beginning with wh-words. This approach would simplify the subcategorization problem as well.

relevant are those that follow verbs and prepositions then the grammar writer can describe the structure of NPs in general, but attempt to build that structure only after a verb or preposition. The fact that the subject NP would then remain unparsed poses no problem for the parsing algorithm.

Beyond the parsing engine, the major resources are the parser rules and the lexical sources. The approximately one-hundred[8] parser rules we used are a subset of a fairly large grammar written by DeMarcken (1990). The subset was selected with the aim of meeting the requirements for structure described in Section 2.2. There is no reason not to use a more complete grammar other than the goal of demonstrating how little syntactic knowledge is needed to use the progressive and rate-adverb tests. The lexical syntactic categories used by the parser are a subset of those defined by the tagged LOB corpus. These include about ten open categories, roughly the ones found in any collegiate dictionary.[9] Even the most basic subcategorization information, such as whether a verb is transitive or intransitive, is unavailable.

In addition to the parsing engine, parsing rules, and dictionary, we used a variation on the DeRose (1988) statistical disambiguator (described in DeMarcken, 1990) to deal with lexical ambiguity. Fortunately, the LOB corpus is also available in a tagged form, where each word has been manually disambiguated. In order to check the effects of errors in statistical disambiguation on our conclusions we compared our results using the disambiguator to the results we got using the hand-tagged corpus. Specifically, we compared the number of progressive and rate-adverb constructions found for each of the verbs in the corpus. To our great surprise, the results were absolutely identical whether we used the disambiguator or the hand-tagged corpus. Although our disambiguator is fairly accurate, this identical performance on 100,000 verb occurrences dramatically demonstrates how insensitive the identification of these constructions is to lexical syntactic error. This one observation is perhaps the most impressive of all our results so far in demonstrating the ease of using these tests under the adverse circumstances with which children are faced.

3 Scaling Up

In order to realize the promise of the approach we have taken to semantic classification we must scale it up beyond the demonstration described above. Indeed, we are currently scaling it up both in terms of the quantity and type of text it can learn from and in terms of the lexical semantic classes it can learn. The ability to deal with text from diverse sources and in diverse formats

[8] It is impossible to convey precisely the power of the rules. They are more compact than context-free rules, but they seem not to be greatly so.

[9] Nominally, the LOB uses 132 categories, but most of these are either closed categories (such as all inflections of *do*, *be*, etc.) or inflections of open categories. We ignore the inflections, using our own morphological recognizer instead.

will have both psychological and technological benefits. On the psychological side, we hope to process corpora of transcribed maternal speech as well as the written text we have used so far. That will allow us to model much more accurately the challenges and opportunities that children would face in doing semantic classification from syntactic context. On the technological side, the log-normal frequency distribution of words in text[10] requires us to process very large quantities of text in order to compile a substantial lexicon. The sources of text that are available to us on the hundred-million words scale are completely raw, so a great deal of work is needed merely to divide the text into sentences. In addition we need sources for the major categories of the words, and these will be unavailable for many words, especially proper names. Indeed, even identifying the boundaries of proper names requires work. Despite the technological hurdles, we hope to be learning from newspaper text sometime this year.

The other way in which we must scale up is in terms of the number and diversity of semantic categories in which we can classify verbs. We are currently working on telicity, the distinction between processes and accomplishments/achievements (Brent, 1989). Work is also underway on the distinction between stage-level and object-level stative predicates (Dowty, 1979). Many of the semantic categorizations that we hope to learn, including telicity, are properties of the subcategorization frame in which a verb appears as well as of the verb itself. This means that our parser needs to be able to determine subcategorization frames. Although it currently goes a long way toward doing so, several hurdles remain. Nonetheless, as our ability to determine subcategorization improves we hope to be able to classify verbs of locomotion by their strong tendency to occur with directional prepositions and no direct object. Indeed, there are many classifications that can be learned using this type of subcategorization information. By expanding in this direction, we hope to make contact with the data on child language acquisition collected by Gleitman (1989), Pinker (1989), and others.

4 Conclusions

The data presented in this paper suggest that, with very modest syntactic and lexical capabilities and no semantic knowledge, it is possible to exploit the progressive and rate-adverb constructions to readily determine the availability of non-stative senses for many verbs. This conclusion is significant both for the study of child language acquisition and for the technology of automated, scientific lexicography. We do not claim that child learners exploit these particular constructions to do semantic classification. Rather, the psychological significance of this work lies in its demonstration that the syntactic contexts of verbs provide reliable information about their semantics which can be recovered with

[10] Zipf (1949)

minimal sophistication. The significance of our conclusion for scientific lexicography is its promise of permitting the automatic construction of large-scale lexica from primary sources. The automatic construction of such lexica is, in turn, one of the most promising paths in natural language processing technology.

References

[Boguraev and Briscoe, 1987] B. Boguraev and T. Briscoe. Large Lexicons for Natural Language Processing: Utilising the Grammar Coding System of LDOCE. *Comp. Ling.*, 13(3), 1987.

[Brent, 1989] M. Brent. Earning dividends on lexical knowledge: How the rich can get richer. In *Proceedings of the First Annual Workshop on Lexical Acquisition*. IJCAI, 1989.

[Brown, 1973] R. Brown. *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA, 1973.

[Church and Hanks, 1990] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, 1990.

[DeMarcken, 1990] C. DeMarcken. Parsing the LOB Corpus. In *Proceedings of the Association for Computational Linguistics*. Assocation for Computational Linguistics, 1990.

[DeRose, 1988] S. DeRose. Grammatical category disambiguation by statistical optimization. *Comp. Ling.*, 14(1), 1988.

[Dowty, 1979] D. Dowty. *Word Meaning and Montague Grammar*. Synthese Language Library. D. Reidel, Boston, 1979.

[Gleitman, 1989] L. Gleitman. The structural sources of verb meanings. Keynote address at Child Language Conference, 1989.

[Knight, 1989] K. Knight. Integrating language acquisition and knowledge acquisition. In *Proceedings of the First Annual Workshop on Lexical Acquisition*. IJCAI, 1989.

[Lakoff, 1965] G. Lakoff. *On the Nature of Syntactic Irregularity*. PhD thesis, Indiana University, 1965. Published by Holt, Rinhard, and Winston as *Irregularity in Syntax*, 1970.

[Landau and Gleitman, 1985] B. Landau and L. Gleitman. *Language and Experience*. Harvard University Press, Cambridge, MA, 1985.

[Pinker, 1984] S. Pinker. *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA, 1984.

[Pinker, 1989] S. Pinker. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA, 1989.

[Zernik and Dyer, 1987] U. Zernik and M. Dyer. The self-extending phrasal lexicon. *Comp. Ling.*, 13(3), 1987.

[Zipf, 1949] G. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, New York, NY, 1949.