Part of Speech Induction from Distributional Features: Balancing Vocabulary and Context

Vivek V. Datla
Dept. of Computer Science and
Institute for Intelligent Systems,
Univ. of Memphis
365 Innovation Drive, Memphis,
TN 38152, USA

King-Ip Lin
Dept. of Computer Science,
Univ. of Memphis
Dunn Hall, Memphis,
TN 38152, USA

Max M. Louwerse
Dept. of Psychology and
Ins. for Intelligent Sys., Univ. of Memphis
365 Innovation Drive Memphis, TN 38152, USA
Tilburg Centre for Cognition and Comm. (TiCC),
Tilburg Univ., PO Box 90153, 5000 LE,Tilburg, Netherlands.

Abstract

Past research on grammar induction has found promising results in predicting parts-of-speech from n-grams using a fixed vocabulary and a fixed context. In this study, we investigated grammar induction whereby we varied vocabulary size and context size. Results indicated that as context increased for a fixed vocabulary, overall accuracy initially increased but then leveled off. Importantly, this increase in accuracy did not occur at the same rate across all syntactic categories. We also address the dynamic relation between context and vocabulary in terms of grammar induction in an unsupervised methodology. We formulate a model that represents a relationship between vocabulary and context for grammar induction. Our results concur with what has been called the word spurt phenomenon in the child language acquisition literature.

Introduction

Understanding language requires an understanding of the structure of the language. Such syntactic knowledge is critical for humans and computers alike. For humans grammar induction, however, seems straightforward, with any child exposed to language acquiring its syntax seemingly effortlessly. For natural language processing this is far less straightforward.

There are two common computational strategies used to identify syntactic elements of language, they are top-down and bottom-up strategies. Top-down (or rule-based) strategies include syntactic parsers which use predefined grammatical rules. The task is to map the lexicon of the target language in these predetermined syntactic categories. This process requires a large set of rules to map the word to an existing syntax. Part of speech induction is typically done in such a top-down fashion (Brill 1995; Garside and Smith 1997; DE Haan 2000).

Top-down strategies use some predefined rules, such as deciding on the number of hidden states in a hidden markov model, or making simple rules such as identifying words ending in *ly* as adverbs etc. Different rules apply for different languages, and even though the rules are rigid, the performance of top-down syntactic parsers is high (Brill 1995; Garside and Smith 1997; DE Haan 2000).

Bottom-up strategies for syntactic parsing use unconstrained mechanisms in which there are no predefined categories to build the syntax. Instead, the process depends on patterns that are found in the incoming data in order to induce category membership. Without rules being present to reduce the number of possibilities, as is the case for top-down parsing, rules need to be generated from the data without any external supervision (Redington, Chater, and Finch 1998).

In order to generate grammar rules, the fundamental idea is to capture the relations among the words, such as their co-occurrence and their frequency distribution with respect to other words in the vocabulary.

Various studies have investigated the effect of bootstrapping syntax from frequently used English words, including (Redington, Chater, and Finch 1998; 1993; Finch and Chater 1992). Their work showed that distributional information can be used successfully to induce parts of speech. Taking the 1000 most frequent words from the CHILDES corpus (MacWhinney 1995) and 150 words as context words from their vocabulary, they were able to bootstrap syntactic categories with decent accuracy of around 75%.

Redington, Chater, and Finch (1998); Redington, Chater, and Finch (1993); Finch and Chater (1992) attempted to induce grammar using distributional features but they gave fixed amount of context information to induce different kinds of grammatical items, and allowed different grammatical items to cluster together.

Even though their results are impressive, it is important

to explore to what extent a more natural lexical environment with varying amounts of context information affects their results, and to what extent standard syntactic categories (not clusters of these categories) can be obtained in the grammar induction task.

In the current paper we thus extended the work in (Redington, Chater, and Finch 1998) by looking into the amount of context information that is needed to induce the categories of the parts-of-speech. We also formulated the dynamic relation between context, vocabulary and purity of the grammatical items induced by distributional features.

Method

Our goal was to organize the words in the vocabulary into various syntactic groups, with each group containing words that are related syntactically, having the same part-of-speech.

In order to achieve this, we needed a vocabulary (the distinct words in a language) and information about linguistic contexts in which a word occurs. In order to bootstrap the syntax from the most common words in English, we used the top 1000 words from the Web 1T 5-gram corpus (Brants and Franz 2006) as our vocabulary. The Web 1T 5-gram corpus consists of one trillion word tokens (13,588,391 word types) from 95,119,665,584 sentences. The volume of the corpus allows for an extensive analysis of patterns in the English language. We made the assumption that these 1000 words represent the total vocabulary in our language.

We represented distributional information of a word in the form of a vector, its distributional vector. In order to find the distributional vector that represents the relation between vocabulary word word1 and context word word2 we calculated the frequency of the linguistic context right after the vocabulary word like {word1 word2}, one word after the vocabulary word like {word1 word0 word2}, context occurring before the vocabulary word like {word2 word1}, and context occurring before one word before the vocabulary word like {word2 word0 word1}. Because of the four possible locations of the context word with respect to the vocabulary word, each vocabulary word and context word yielded four values.

In order to find the similarity between the distributional vectors we used Spearman Rank correlation, following Redington, Chater, and Finch (1998). We next created an n-by-n similarity matrix where n is the size of the vocabulary. We next used agglomerative hierarchical clustering (Manning, Raghavan, and Schütze 2008) to merge instances into clusters, using average link as the distance measure, following Finch and Chater (1992); Redington, Chater, and Finch (1998). Average link distance is the distance between two clusters set by the averages of all similarities between the instances in the cluster.

Analogous to linguistic categories, where a word belongs to one syntactic category only (perhaps with the exception of homographs), we measured the purity of each cluster in a very strict sense, allowing each cluster to belong to only one category of parts of speech. Importantly, since we do not know the label of each cluster, we took a greedy labeling approach. That is, if the majority of the words inside the cluster belonged to a particular parts of speech then we labeled the cluster with the same parts of speech.

Purity is a simple and logical measure, as it looks at the accuracy of clustering at a local level, at each cluster, and does not affect the overall accuracy of multiple clusters of the same grammatical items. In order to measure purity of the cluster we took a simple ratio, namely the number of words belonging to the same parts of speech as the cluster label to the total number of words present in the group.

To determine the effect of vocabulary size on the performance of the grammar induction, we collected 300 words randomly from the 1000 word vocabulary, then we collected 400 words randomly, then 500 words, until we used the entire set, resulting in eight sets of vocabulary. For context size we randomly picked 10, 50, 100, 150, 200, 250, 300, 400, 500, resulting in 9 sets of context.

We ran each experiment 100 times for every combination context and vocabulary size. In each run we randomly chose the words for the required combination, thus reducing the selection bias of the context and vocabulary words. The results reported in this paper indicate the averaged accuracy over the 100 runs.

To establish a baseline in order to measure the purity of our clusters we used the standard POS tagger CLAWS (Garside and Smith 1997). CLAWS is a hybrid tagger which takes advantage from the HMM models and probabilistic methods. The CLAWS hybrid tagger tags words into 137 categories, for simplicity we have merged the similar grammatical items.

For example 23 sub categories of nouns such as proper noun, common noun, numeral noun etc. were merged to one category of noun. Similar process was employed on all the other grammatical items and 137 categories are merged to form 11 categories of grammatical items.

The 11 grammatical categories obtained after the merging process are as follows: ADJ (adjective), ADV (adverb), ART (article), CONJ (conjunction), wDET (word before/after determiner), N (noun), NUM (number), PREP (preposition), PRON (pronoun), V (verb), and alphabet (used for symbols).

We have used CLAWS as our gold standard as it has demonstrated having a better accuracy 95% (Van Rooy and Schäfer 2003) than the Brill (Brill 1995) and the TOSCA-ICLE tagger (DE Haan 2000). An alternative to CLAWS might seem a lexical and syntactic database like CELEX (Baayen, Piepenbrock, and Gulikers 1995). However, the nature of the Web 1T 5-gram corpus is such that the top 1000 words from this corpus are not present in the database.

Table 1 shows the instances of our vocabulary based on the types of POS as found by the CLAWS tagger. It is not surprising to see more than 50% of the words in the dataset to be nouns. The majority categories of the words are nouns (59%), verbs, adjectives, adverbs, pronouns, and prepositions.

The ratio of number of determiners, pronouns, adverbs, prepositions and conjunctions to the total number of words decreases as vocabulary increases. This decrease is expected as they are closed set of items that occur frequently in language.

To summarize we developed distributional feature vectors for each word in the vocabulary based on the context. Similarity between these vectors is used as a distance measure to cluster the words into groups using agglomerative clustering. We then check the purity of each cluster in a strict sense, forc-

Table 1: Vocabulary sizes and distribution among various categories

Vocabulary Size	N	V	ADJ	ADV	PRON	ART	wDET	PREP	CONJ	A	NUM
300	0.4	0.2	0.06	0.10	0.05	0.01	0.05	0.06	0.03	0.01	0.02
400	0.44	0.19	0.09	0.09	0.04	0.01	0.04	0.05	0.03	0.01	0.01
500	0.49	0.18	0.09	0.08	0.04	0.01	0.03	0.04	0.02	0.01	0.01
600	0.53	0.19	0.08	0.07	0.03	0.01	0.03	0.04	0.02	0.01	0.01
700	0.54	0.18	0.09	0.06	0.03	0.01	0.02	0.03	0.02	0.01	0.01
800	0.56	0.18	0.09	0.06	0.02	0.01	0.02	0.03	0.02	0.01	0.01
900	0.58	0.17	0.09	0.06	0.02	0.00	0.02	0.03	0.02	0.01	0.01
1000	0.59	0.17	0.09	0.05	0.02	0.00	0.02	0.03	0.01	0.01	0.01

ing each cluster to belong to exactly one category of parts of speech. The label given to each cluster is decided on the majority of the word category inside that cluster.

Results

Figure 1 shows the accuracy of the induced grammatical items in relation with vocabulary and context. The vocabulary varies from 300 to 1000 words, and the context varies from 10 words to 500 words.

For nouns one can observe that very little context was needed to cluster them. Even with a context as little as 10 words we could cluster them with an accuracy greater than 0.94, indicating that nouns cluster easily even with small context information.

For all grammatical items we observe that there is a sharp increase in accuracy initially as the context increases, for a fixed vocabulary and then the accuracy plateaus.

We also observe that as vocabulary increases for a fixed context, the accuracy of the syntax falls. This can be attributed to the relative decrease in the power of distributional information to differentiate various new words that are being added into the vocabulary. This is observed clearly in all the categories of the parts of speech except for nouns. For nouns we do not see this effect as words obtain their true syntactic category with very less distributional information.

We have shown support for the idea that syntactic category acquisition is dependent on context, and the accuracy is comparable to the top down approaches. This indicates that grammatical categories of the words can be inferred by having relative information among the words.

We also show that all the grammatical items are not created equal. Nouns are the easiest to learn as they require very little context, adjectives (figure 1(a)) and adverbs (figure 1(c)) are difficult to learn and bootstrapping methods have also given 40% to 60% accuracy.

The categories of prepositions (figure 1(d)), verbs (figure 1(b)), adverbs (figure 1(c)), and pronouns (figure 1(f)) also showed a similar pattern of increasing accuracy when context increases, and then accuracy plateaus after a certain context size. The accuracy is not high when compared to nouns. The categories of adjectives, pronouns, verbs, and adverbs showed a sudden decrease in accuracy when the vocabulary is increased for smaller sizes of context.

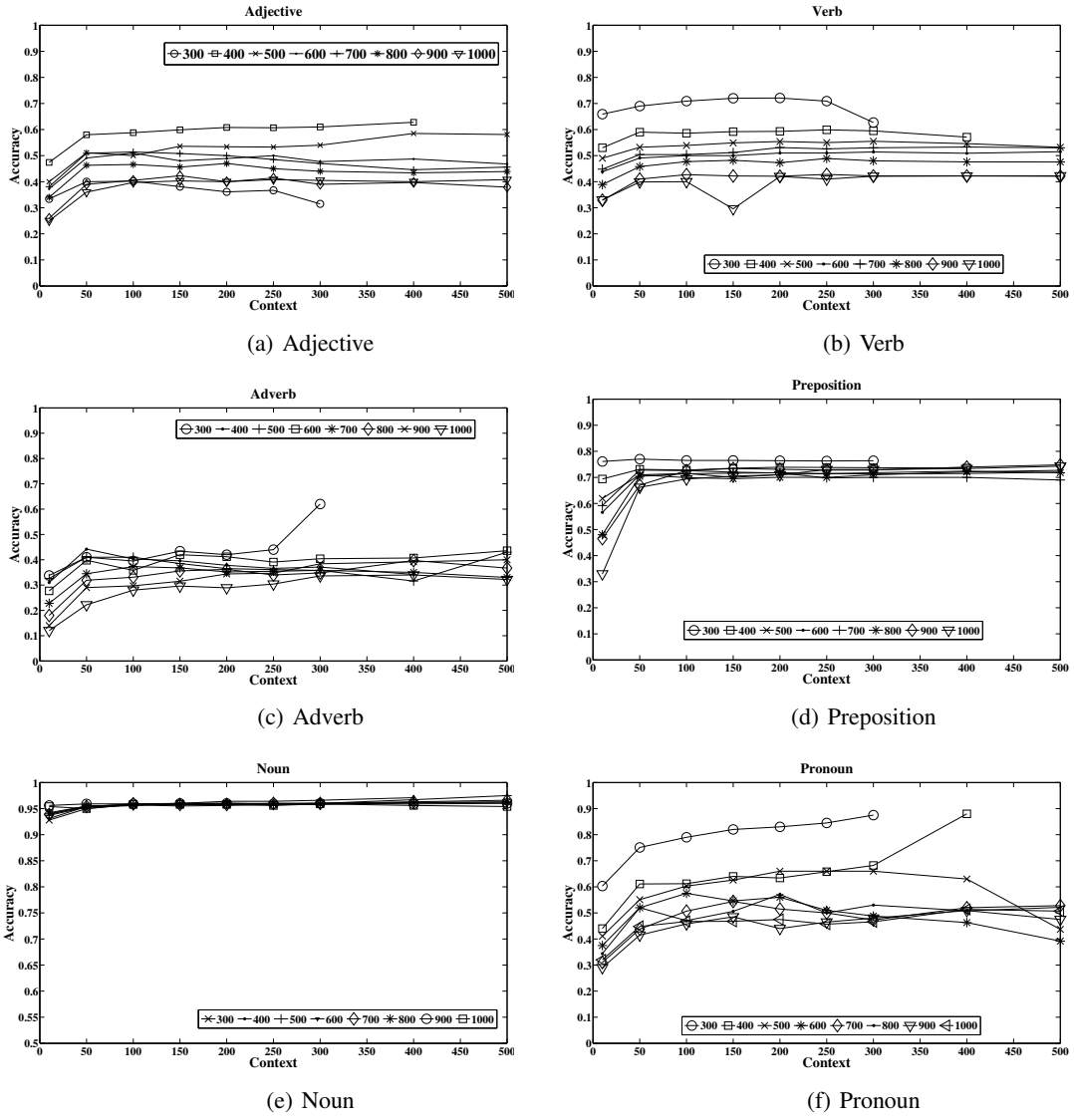
Next we investigate the question, what is the dynamic relationship between vocabulary and context? In order to answer this question we formulated a relationship between context and vocabulary with respect to the accuracy in inducing grammatical items. Figure 2 represents, the overall accuracy of the part-of-speech, where we fit a curve over the various accuracy values obtained for respective context and vocabulary. The curve, which fit the data best, is exponential in nature. The equation that best captures the relationship is $F(x, y) = exp(a * x^c + b * y^d)$ where x = context; y = vocabulary; a = -0.896; b = 0.662; = -0.018; d = -0.0340, the goodness of fit measures for the function are of as follows: SSE = 0.00086, R-square = 0.975, Adjusted R-square = 0.9, RMSE = 0.0033. This curve shows that, overall accuracy saturates quickly even with limited context of around 100 words. The overall accuracy saturates at around 75% even when the context becomes the same size of vocabulary.

Language Acquisition

Understanding a language requires understanding structure of the language. This learning of structure comes easily for humans. A child typically learns around 60,000 words from birth to adulthood, averaging around 8 to 10 words in a day. During the second year of the child the learning of new words increases tremendously, and this phenomenon has been referred to as word spurt, or vocabulary explosion (Gopnik and Meltzoff 1987). Researchers who argued against the word-spurt theory have pointed out that the sudden ability to learn new words to other factors of the child development, such as mental development, memory development, and more experiences the child has as one grows (Bloom 2002).

Researchers who argued for a word spurt (Gopnik and Meltzoff 1987; Gerken, Wilson, and Lewis 2005; Nazzi and Bertoncini 2003; McMurray 2007) indicate that the acceleration is guaranteed in any system which have (1) the words are acquired in parallel, and representations build simultaneously for multiple words, and (2) the difficulty of learning words follows Zipfś law. This line of argument was also given by Louwerse and Ventura (2005) where they showed that even a simple semantic model such as Latent Semantic Analysis (LSA) shows exponential increase in the semantic information it captures as the words increase in the vocabulary, indicating for a presence of word-spurt due to the bootstrapping of relative information gain from the new words acquired.

We are not in a position to resolve the word-spurt debate. However, the results from the bottom-up approach presented here are consistent with the language acquisition literature



(a) Adjective (b) Verb

(c) Adverb (d) Preposition

(e) Noun (f) Pronoun

Figure 1: Accuracy for each part-of-speech category

(Gerken, Wilson, and Lewis 2005; Gopnik and Meltzoff 1987; Louwerse and Ventura 2005; McMurray 2007; Mitchell and McMurray 2008; Nazzi and Bertoncini 2003). Child language acquisition points out that children learn few kinds of nouns easily, our results indicate that children acquire nouns easily, and the nouns also get clustered together as they share strong correlations in terms of distributional information. Prepositions, adjectives, and pronouns also show the pattern that when context increases, the accuracy also increases for a fixed vocabulary. For a fixed context when vocabulary increases the accuracy also decreases for these categories.

Conclusion

In this paper we investigated the relation between context and vocabulary for part-of-speech induction. Specifically we investigated the effect of context and vocabulary size on the accuracy of part of speech categories. The results indicate that the relation between the vocabulary and context is logarithmic, indicating that accuracy improves exponentially when context is less, but plateaus as context increases.

In previous research 150 words has been used as the size of the context (Redington, Chater, and Finch 1998; Schütze 1995), our results demonstrate that categories like nouns, prepositions, adverb, and verb saturate quickly in terms of accuracy for 100 words of context, even for increasing vocabulary. With in the categories we see that nouns have a high accuracy and cluster together even for very small amount of context.

We also show the dynamic relation between vocabulary and context, see Figure 2. For increasing context for a fixed vocabulary we see the accuracy increasing exponentially and then plateauing, and in the other way for a fixed context and

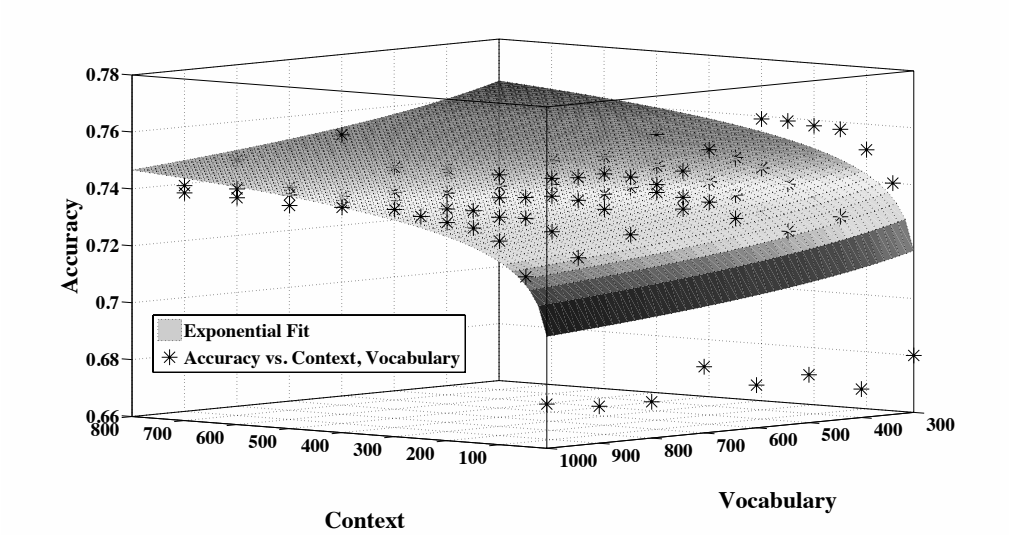


Figure 2: Overall Accuracy as a function of vocabulary and context

increasing vocabulary we see there is a dip in accuracy. The relative information among the words plays an important role in grammar induction.

References

Baayen, H. R.; Piepenbrock, R.; and Gulikers, L. 1995. *The CELEX lexical database. release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Bloom, P. 2002. *How children learn the meaning of words*. Cambridge, MA: The MIT Press.

Brants, T., and Franz, A. 2006. *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.

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.

DE Haan, P. 2000. Tagging non-native english with the tosca-icle tagger. *Language and computers* 33:69–80.

Finch, S., and Chater, N. 1992. Bootstrapping syntactic categories using statistical methods. In *Background and Experiments in Machine Learning of Natural Language*, 229–235. Tilburg University: Institute for Language Technology and AI.

Garside, R., and Smith, N. 1997. A hybrid grammatical tagger: CLAWS4. In Garside, R.; Leech, G.; and McEnery, T., eds., *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London, UK: Longman. 102–121.

Gerken, L.; Wilson, R.; and Lewis, W. 2005. Infants can use distributional cues to form syntactic categories. *Journal of child language* 32(2):249–268.

Gopnik, A., and Meltzoff, A. 1987. The Development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development* 58(6):1523–1531.

Louwerse, M. M., and Ventura, M. 2005. How children learn the meaning of words and how lsa does it (too). *The Journal of the Learning Sciences* 14(2):301–309.

MacWhinney, B. 1995. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 2nd edition.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

McMurray, B. 2007. Defusing the childhood vocabulary explosion. *Science* 317(5838):631–631.

Mitchell, C. C., and McMurray, B. 2008. A stochastic model for the vocabulary explosion. In Love, B.; McRae, K.; and Sloutsky, V. M., eds., *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1919–1926. Austin, TX: Cognitive Science Society.

Nazzi, T., and Bertoncini, J. 2003. Before and after the vocabulary spurt: two modes of word acquisition? *Developmental Science* 6(2):136–142.

Redington, M.; Chater, N.; and Finch, S. 1993. Distributional information and the acquisition of linguistic categories: A statistical spproach. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 64–70. Austin, TX: Cognitive Science Society.

Redington, M.; Chater, N.; and Finch, S. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 22(4):425–469.

Schütze, H. 1995. Distributional part-of-speech tagging. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, EACL '95, 141–148. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Van Rooy, B., and Schäfer, L. 2003. An evaluation of three pos taggers for the tagging of the tswana learner english corpus. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, 835–844.