Parsing Using a Grammar of Word Association Vectors

Robert John Freeman(✉)

EasyChair, Sydney, Australia
rob@chaoticlanguage.com

Abstract. This paper was first drafted in 2001 as a formalization of the system described in U.S. patent U.S. 7,392,174. It describes a system for implementing a parser based on a kind of cross-product over vectors of contextually similar words. It is being published now in response to nascent interest in vector combination models of syntax and semantics. The method used aggressive substitution of contextually similar words and word groups to enable product vectors to stay in the same space as their operands and make entire sentences comparable syntactically, and potentially semantically. The vectors generated had sufficient representational strength to generate parse trees at least comparable with contemporary symbolic parsers.

1 Motivation

1.1 Basic Intuition

The basic intuition underlying our formalisation is that a word can be characterised by enumerating words that occur in similar contexts, e.g. Menno van Zaanen [21] attributes to Harris [4] the "notion of interchangeability":

> constituents of the same type can be replaced by each other

There are variations on this (c.f. if constituents are not of the same type they involve a change of meaning and cannot be replaced), but it involves a comparison, and we will refer to it in general as the "contrastive method".

This intuition is implicitly present in almost all linguistic classifications. E.g. the definition of a part of speech such as *determiners* is based on the observation that there is a group of words that exhibit similar behaviour in a range of contexts:

- preceeding {"cat", "dog", "car", "wardrobe"...}
- preceeding {"big", "small", "blue", "bright"...}
- following {"take", "put", "see"...}
- ...

Of course, in the traditional way of classifying language units such observations are followed by very strong generalisations. But is this reasonable?

B. Goertzel et al. (Eds.): AGI 2021, LNAI 13154, pp. 86–95, 2022.
https://doi.org/10.1007/978-3-030-93758-4_10

1.2 Idiosyncratic Usage

A strong theme in linguistics for many years has been the idiosyncrasy of language use defying classification. For an example which appears to date to Halliday [6], "strong" and "powerful" can be used interchangeably in many contexts, but not in the context of "tea", viz. "strong tea"/"*powerful tea". Pawley and Syder [5] present long lists of examples which defy codification in rules under the title of "native-like selection", which they characterize as a "puzzle" for linguistics.

In fact the Applied Linguistics literature is full of observations that language appears to have no rules without exceptions, and indeed in some senses to consist of exceptions only. Nattinger [18], Weinert [20], and Lewis [17] provide a sampling, reaching an extreme with Lewis's "Lexical Approach".

This is addressed less often in a machine learning context, but even here Dagan, Marcus, Markovitch note, Dagan [3]:

> It has been traditionally assumed that ... information about words should be generalized using word classes... However, it was never clearly shown that unrestricted language is indeed structured in accordance with this assumption.

1.3 The Phoneme

But we can go further than the lack of proof noted by Dagan et al. It was shown not to be so. In fact, you could characterize this is as the central dilemma of linguistics for the last 50 years. It is behind the major divisions which have characterized the subject during that time. Put more strongly, what we find when we try to abstract language structure from observation is that it contradicts itself. And this was shown to be so a long time ago for the first category, and great success of the contrastive method, the phoneme.

The crisis was catalysed by the polarizing figure of 20th century linguistic theory, Noam Chomsky. Chomsky brought a fresh perspective to many issues in linguistics, but while it has been largely forgotten it is fair to say his attack on the phoneme was one of the most important at the time. E.g. Hill [15]:

> I could stay with the Transformationalists pretty well, until they attacked my darling, the phoneme.

Contrastive methods along the lines attributed to Harris above were the standard method of linguistic analysis. The phoneme was the major success of the method. Chomsky pointed out that these contrastive methods led to contradictory results, and linguistics shattered. As Newmeyer [19] says while discussing Chomsky's Logical Basis of Linguistic Theory [1]

> Part of the discussion of phonology in 'LBLT' is directed towards showing that the conditions that were supposed to define a phonemic representation (including complementary distribution, locally determined biuniqueness, linearity, etc.) were inconsistent or incoherent in some cases and led to (or at least allowed) absurd analyses in others.

So abstracting language categories from observation led to "inconsistent or incoherent ... analyses". Or equivalently that elements of language combine non-linearly. Lamb [12] says of Chomsky:

> He correctly pointed out that the usual solution incorporates a loss of generality, but he misdiagnosed the problem. The problem was the criterion of linearity.

See also Lamb 1966 [9], Lamb 1967 [8], Lamb 1976 [10], Lamb 1998 [11].

1.4 The (Very) Short History of Machine Learning in Theoretical Linguistics

This was shown to be the case for the phoneme, but it was not seriously contested for other categories. The contrastive method never got beyond the phoneme. Instead linguistics moved beyond abstraction of categories by contrastive analysis, fragmenting into schools which on the one hand continued to respect contrastive analysis but rejected structure as a meaningful parameter, like Functionalism, or on the other hand continued to respect language structure, but rejected the idea it could be learned from observations, Generativism. Generativism hypothesized that the contradictions observed when you tried to abstract categories from observations were not relevant to the system of language at all. It was still assumed structure was relevant, and objective, but because learnable categories had been observed to contradict (among other reasons), structure was assumed to be innate.

Unfortunately, having split the field, this key issue of contradictions in learnable categories was largely forgotten and unavailable to those outside linguistics seeking insights. So when rapid developments in computing made large amounts of data available, non-linguists started trying to learn from it. Non-linguists in the main, because to reiterate, theoretical linguistics had already split into schools which assumed language structure was unlearnable, or irrelevant, making the issue disappear. It is ironic that for this very reason the non-linguists perhaps felt linguistics was irrelevant and ignored it, in its turn, because it did not address the issue which interested them viz. learning categories from observable data.

1.5 Contradiction in Language Categories

This then is our motivation for seeking to model grammar as a vector product of word similarity vectors, most forcefully that abstractions from observations were shown to lead to contradictions, or in the case of Lamb "non-linearity". That this is so is actually a rare point of agreement in theoretical linguistics, it is only interpretations which have varied: Universal Grammar, Functionalism, Cognitivism.

Unfortunately for a rare point of agreement in linguistics, it is widely ignored. Notably it appears to be ignored in the field of machine learning. It might be seen to render the entire field of grammatical induction moot.

Chomsky concluded that if a coherent language structure could not be learned from observations, then it must be innate. Functionalists (and various schools of Cognitivism) concluded it was irrelevant. But unlearnable need not mean innate or irrelevant. It may simply mean generalizations can only be ad-hoc. We have a better understanding of complexity in formal systems now. Unlearnable, random (in a specific technical sense e.g. Kolmogorov [7], Chaitin [2]), or even chaotic, systems are starting to seem more like the norm. As far as the author is aware these kinds of complexity issues were first suggested, at least in the context of computational solutions, in an earlier discussion of the motivations surrounding this parser in Freeman 2000 [14]. But it is possible to find similar ideas in linguistics, notably Hopper's Emergent Grammar 1987 [16]:

> The notion of emergence is a pregnant one. It is not intended to be a standard sense of origins or genealogy, not a historical question of 'how' the grammar came to be the way it 'is', but instead it takes the adjective emergent seriously as a continual movement towards structure, a postponement or 'deferral' of structure, a view of structure as always provisional, always negotiable, and in fact as epiphenomenal, that is at least as much an effect as a cause.

Possibly Lamb's "non-linearity" is the same.

As suggested in that 2000 paper, this ad-hoc character should by no means be regarded as a bad thing. Rather it hints at a far greater richness of structure available to language.

2 Formalization

According to the method of this paper then, in order to prevent relevant information from being lost in the task of modelling natural language, we need a concept of overlapping classes that can be constructed, ad-hoc, on the fly as opposed to having to be pre-defined.

2.1 Similarity

Thus we stop at the primary observation that "word w is similar to the words $w_1, ..., w_k$ because it appears in similar contexts" and treat $w, w_1, ..., w_k$ as class label for w.

To put this more formally, we need to define two things: what is actually a context and what *similar contexts* means. A simple initial approximation to the former question is to define the context of a word (or word group) w as the set of all word pairs w', w'' such that there is at least one occurence of the triplet w', w, w'' in the available corpus.

$$Con(w) := \{(w', w'') : \exists i : X_i = w', X_{i+1} = w, X_{i+2} = w''\}$$

Now the similarity of two words w and w' can be measured by the percentage of common contexts relative to all contexts they appear in. Dekang Lin provides a principled measure [13]:

$$sim(w, w') := \frac{2 \times I(Con(w) \cap Con(w'))}{I(Con(w)) + I(Con(w'))}$$

Where I is the information.

With the two definitions formulated above, the class a word belongs to can be defined as a set of similarity values between words as below. For convenience we limit the set to include only words similar beyond a threshold C:

$$class(w) := \{w' : sim(w, w') > C\}$$

e.g. A segment from the entry from "somewhat" estimated for one of our corpora is:

```
somewhat {somewhat,1 significantly,0.047 slightly,0.043
considerably,0.036 substantially,0.025 far,0.025 a&lot,0.024
more&and,0.024 one&or,0.024 much,0.022 becoming,0.021
likely&to&be,0.021 a&little,0.02 considered,0.018 nothing,0.017
rather,0.017 so&much,0.016 relatively,0.016 something,0.015
certainly,0.014 quite,0.014 generally,0.014 getting,0.014 still,0.012
therefore,0.011 done,0.011 often,0.011 of&course,0.01 even,0.0097
very,0.0083 always,0.0082 a&very,0.008 once,0.0073
further,0.0072 little,0.0071 less,0.007 really,0.0066...}
```

Where the vector is made up of a set of (word, score) pairs, and the score is the value of the chosen similarity function between the pair "somewhat" and the other words in the vocabulary.

For practical reasons it is more convenient to represent the class of word w as a vector of real numbers, each representing the similarity of a vocabulary word to w. So if $w_1, ..., w_N$ is the vocabulary of our language, then the class of w is defined by the following function ϕ:

$$\phi(w) := (\phi_1(w), ..., \phi_N(w))$$

2.2 Application to Trees

We now extend the formalization to estimate representations for potentially unobserved sequences of words. We say "potentially" unobserved, because the sequences may in fact be observed, however we seek to extend the representation in the absence of observation. Essentially we seek to represent syntax, specifying which new combinations of words are acceptable to the language.

We will call these potentially unobserved sequences t for "trees", because the formula is non-associative. Different orders give different products, so the product has structure, like a tree. This can be seen as an advantage of the method, it predicts phrase structure.

In the vector formulation, we have a vocabulary V: $w_1, ..., w_N$ of words and word sequences. Let the function $\delta(i,j) \rightarrow \{0, ..., N\}$ map a pair of elements in V to another entry in V. Simply put, let it map a pair into an entry for the pair together, if they are observed to occur together ($\delta(i,j) = 0$ means that no such entry exists, i.e. i and j are not observed to occur together.)

Formally, the extension of the word class function to binary trees is defined recursively, if t is a word, then $\phi_i(t)$ reduces to $sim(t, w_i)$, if t is a binary tree consisting of subtrees t_1 and t_2, then we define the components of a new vector representing the combination of trees to be:

$$\phi_k(t) := \sum_{i,j} MI(w_i, w_j) \cdot \phi_i(t_1) \cdot \phi_j(t_2) \cdot \phi_k(\delta(i,j))$$

In words: if a pair of words between the components of the operand trees are observed to occur together ($\delta(i,j) \neq 0$), the components for their observed combination ($\phi_k(\delta(i,j))$) contribute to a new component for the combination of the operand trees ($\phi_k(t)$.) Note: MI is the common "mutual information" between two words occurring in sequence, which was used in this case to scale the significance of an observed pairing. We in no way regard this choice as exhaustive or exclusive. In much the same way as the similarity measure described in Sect. 2.1 above, many scalar association measures might be considered.

Essentially we seek to generate product vectors by an aggressive substitution of contextually similar words and word groups (all the $\phi_k(\delta(i,j))$).

2.3 Scoring a "Parse"

It was assumed that at each stage of association the most grammatical grouping would generate the greatest number or greatest concentration of grammatically similar elements, and an order of association, branching, or the best phrase structure tree, was selected on that basis. Broadly speaking, for each order of association of words t the greatest of:

$$max\{\sum_k \phi_k(t)\}$$

It can be seen as a kind of energy maximum clustering for the path of branching which resulted in the largest substitution groupings.

Selection of a pair as a parse was not based on pairwise co-occurrence of the pair itself, but looking outside the pair, at its context, and counting the number of substitution groupings which could be made for that pair, in that context. It was based on the extent to which a pair could be substituted for by other sequences, as observed in a corpus, and especially the extent to which it could be substituted for by single words.

What was new to the method was not the use of such substitution groupings, however. Such substitution groupings can actually be seen as somewhat standard in machine learning of grammar, indeed dating back to the "notion of interchangeability" attributed to Harris, earlier, which was applied with such early success to the phoneme. What was different with the vector parser was the assumption that such groupings needed to be generated anew for each new sentence presented to the system. That such groupings might be an expansion rather than a compression of structure in a corpus.

It is that such substitution groupings might need to be seen as an expansion, rather than a compression as hitherto assumed in the history of machine learning, which is the central conjecture of this paper.

The vector parser was an early attempt at an implementation of that central conjecture.

More detail, and actual segments of early code, can be found in the technical specification for patent US7392174B2.

3 Discussion

The system described in Sect. 2 was the vector parser system as originally formulated between 1998–2001. It was not picked up at the time, although tests were done comparing performance quantitatively for the task of parsing Chinese in 2003. We hope to present these results in a later paper. In general the parser performed on a level comparable with a contemporary symbolic parser, but not greatly better (see Appendix for a sample of results.)

It now seems apparent that while classes are not abstracted until runtime in this formulation, important details are lost when word similarity is assessed. Not only class, but similarity between words can, and typically does, vary between contexts. What is needed are vectors of contexts, not similarity based on context.

This should be a fruitful direction for further investigation.

However the method of generating product vectors by an aggressive substitution of contextually similar words and word groups (all the $\phi_k(\delta(i,j))$ above) was quite successful. It enables product vectors to stay in the same space as their operands and makes entire sentences comparable syntactically, and potentially semantically. The vectors generated had sufficient representational strength to generate parse trees at least comparable with contemporary symbolic parsers.

In the context of contemporary explorations of vector combination for semantic representation, this approach of aggressive substitution of contextually similar words and word groups is recommended as a possible solution to problems which present themselves of expansion of representation space, representation strength and comparability.

Acknowledgements. Many thanks to Wojciech Skut (wojciech@google.com) for the formalism in this paper and more recent comments. His formalization was what made me first realize the "ad-hoc arrangements of examples" model I was proposing could be seen as a kind of vector product.

A Appendix - Evaluation Sample

A.1 Disclaimer

A formal evaluation of the vector parser was carried out for Chinese by Hu Guoping of iFlytek, China in 2003. A sample of results as reported by him are included here, with his permission. They are provided on an as is basis, for information purposes. For fuller detail inquiries should be directed to Hu Guoping.

A.2 Evaluation Methodology and Results

Word similarity vectors were generated from a Chinese corpus sample of 200Mb. Some pre-processing was done for the Chinese text. Because Chinese is not normally segmented into words a dictionary was used to segment the texts. In principle this should not be necessary, but it was an expedient to make Chinese processing as closely comparable as possible to English. Parsing results were compared for several hundred test sentences with a reference tree-bank.

The reference tree-bank was multi-branching.

E.g.

(并 同)(美国总统 克林顿)(正式 进行 会谈)

Because the vector parser attempts to label all pairwise associations it was not possible to score matches exactly. It was decided to score matched phrases as correct if their first and last characters matched, independently of internal structure. In addition the direction of association of the matched phrases was scored. So 并 has a right direction because it is combined with the right word, 同 has a left direction because it is combined with the left word, 进行 has a middle direction: because it is combined with both left and right words.

E.g.

```
Vector parser: (((促进 ((社会 全面) 进步)) 的) 迫切需要)
Tree-bank: (促进 社会(全面 进步)的 迫切需要)
The Result is 1 correct result, 1 error result, 2 Unknown
Vector parser: (致以 (节日 (的 (诚挚 (问候 ((和 良好) 祝愿))))))
Tree-bank: (致以(节日 的(诚挚 问候)和(良好 祝愿)
The Result is 2 correct result, 2 error result, 2 Unknown
Vector parser: (截止到 (去年 (十一月 底)))
Tree-bank: (截止到(去年(十一月 底)))
The Result is 2 correct result, 0 error result, 0 Unknown result
Vector parser: (各 项 (外 汇 (存 款 (已 (达 (二 十 五 (点 (三 (二 亿 美
元)))))))))
Tree-bank: (((各项 外汇)存款)((已 达)(二十五 点 三 二亿)美元))
The Result is 2 correct result, 4 error result, 2 Unknown result
Vector parser: (全行 (国际业务 (((效益 (创 历史)) 最高) 水平)))
```

```
Tree-bank: ((全行 国际业务 效益)(创(历史(最高 水平))))
The Result is 1 correct result, 3 error result, 1 Unknown result
Vector parser: (在 (国际 (银 (坛 (的 (知名度 (和 (声望 (也 越来越
高)))))))))
Tree-bank: ((在(国际(银 坛))的(知名度 和 声望))(也 越来越高))
The Result is 4 correct result, 2 error result, 2 Unknown result
Vector parser: (再次 (超过 (捕捞 产量)))
Tree-bank: (再次(超过(捕捞 产量)))
The Result is 2 correct result, 0 error result, 0 Unknown result
```

The differences in annotation system meant scored precision was decreased, because generally it was impossible to match all branchings, but this was considered adequate for purposes of comparison.

Table 1. Evaluation with Symbolic Parser

Compared Parser	Setting	Precision1	Recall2	Precision2	F-Score2
Symbolic Parser	基于整句解释的语法分析系统 265 Sentence Single POS input	53.39%	53.48%	67.36%	59.62%
	基于整句解释的语法分析系统, 265 test Sentences All Possible POS input	56.74%	54.90%	69.43%	61.31%
Vector Parser	200M Corpus, -li 60, Original Indextable, 265 test Sentences	73.06%	64.30%	63.15%	63.72%
	Corpus, -li 60, 2nd Indextable, 265 test Sentences	72.96%	63.98%	62.84%	63.41%
	200M Corpus, -li 60, 3rt Indextable, 265 test Sentences	72.37%	63.51%	62.37%	62.94%
	200M Corpus, -li 60, 4th Indextable, 265 test Sentences	76.70%	67.30%	66.10%	66.69%

References

1. Chomsky, N.: The logical basis of linguistic theory. In: Lunt, H.G. (ed.) Proceedings of the Ninth International Congress of Linguists. Mouton & Co., The Hague (1964)
2. Chaitin, G.J.: The Unknowable. Springer, Singapore (1999). https://doi.org/10.1023/A:1015151819254
3. Dagan, I., Marcus, S., Markovitch, S.: Contextual word similarity and estimation from sparse data. Comput. Speech Lang. **9**, 123–152 (1995)
4. Harris, Z.: Methods in Structural Linguistics. University of Chicago Press, Chicago (1951)

5. Pawley, A., Syder, F.: Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, L., Schmidt, I.L. (eds.) 1983: Language and Communication, pp. 191–226. Longman, London (1983)
6. Halliday, M.A.K.: Lexis as a linguistic level. In: Bazell, C.E., Catford, J.C., Halliday, M.A.K., Robins, R.H. (eds.) Memory of J.R. Firth, pp. 148–162 (1966)
7. Kolmogorov, A.: On tables of random numbers. Sankhyā Ser. A. **25**, 369–375 (1963). MR 178484
8. Lamb, review of Chomsky .... American Anthropologist 69, pp. 411–415 (1967)
9. Lamb, Prolegomena to a theory of phonology. Language 42.536-573 (1966) (includes analysis of the Russian obstruents question, as well as a more reasonable critique of the criteria of classical phonemics)
10. Lamb and Vanderslice, On thrashing classical phonemics. LACUS Forum 2.154-163 (1976)
11. Lamb, Linguistics to the beat of a different drummer. First Person Singular III. Benjamins, 1998 (reprinted in Language and Reality, Continuum, 2004)
12. Lamb, Funknet, 24th June, 2004
13. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 296–304. Morgan Kaufmann (1998)
14. Freeman, R.J.: Example-based complexity-syntax and semantics as the production of ad-hoc arrangements of examples. In: Proceedings of the ANLP/NAACL 2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems, pp. 47–50
15. Hill, A.A.: How many revolutions can a linguist live through. Studies in the History of Linguistics, vol. 21 (1980)
16. Hopper, P.: Emergent grammar. Berkeley Ling. Conf. (BLS) **13**, 139–157 (1987)
17. Lewis, M.: The lexical approach: The state of ELT and the way forward. Language Teaching Publications, Hove, England (1993)
18. Nattinger, J.: A lexical phrase grammar for ESL. TESOL Q. **14**, 337–344 (1980)
19. Frederick, J.: Newmeyer. Routledge, Generative Linguistics a historical perspective (1996)
20. Weinert, R.: The role of formulaic language in second language acquistion: a review. Appl. Linguis. **16**(1), 180–205 (1995)
21. van Zaanen, M.: Bootstrapping syntax and recursion using alignment-based learning. In: Proceedings of the Seventeenth International Conference on Machine Learning, pp. 1063–1070 (2001)