Limits to annotation precision

Geoffrey Sampson and Anna Babarczy
School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton
BN1 9QH
England
{geoffs, annab}@cogs.susx.ac.uk

Abstract

This paper seeks to draw attention to a large gap in the current spectrum of research relating to treebanks (linguistically interpreted corpora). There has been substantial work on defining schemes of grammatical annotation, developing automatic parsers for applying those schemes to language samples, and measuring the performance of automatic parsers, but there has been next to no work on establishing the baseline of human analytic accuracy – although this is logically necessary as an underpinning to the other three activities. Apart from arguing this case, the paper also briefly describes a forthcoming pilot project which will aim to begin to fill the gap it identifies.

0. Introduction

Treebanks consist of language samples annotated with structural information – centrally, the grammatical structure of the samples, though some resources include categories of information other than "grammar" sensu stricto. The data contained in treebanks are useful for diverse theoretical and practical purposes. But if we consider not the applications of treebanks but the activities involved in developing them, most of these activities relate to three headings:

1. systems for deriving structure automatically from unannotated language samples – parsers

2. specification of schemes of annotation – targets for parser output

3. metrics for quantifying parsing accuracy

We shall survey these three areas in turn, before arguing that there ought to be a fourth area.

1. Parsers

A great deal of work was put into automatic parser technology well before treebanks of real-life language began to be developed; much early parser research focused on testbeds consisting of invented examples. One survey of work up to that date is Spärck Jones and Wilks (1983). Automatic parsing was seen as a key technology for many potential natural language processing applications – Obermeier (1989: 69) described it as "[t]he central problem" for virtually all such applications.

Since the advent of treebanks, these have been seen as an important resource for developing better automatic parsers (e.g. Charniak, 1996). Conversely, automatic parsing (commonly with manual intervention) has been used to generate larger treebanks.

2. Annotation schemes

Automatic parsing logically presupposes that we have a concept of what correct structural annotation of particular examples should look like – otherwise a parser could produce any random output and there would be no basis for criticizing it. Perhaps surprisingly, little attention seems to have been paid to defining target parsing schemes in the early decades of research on automatic parsing. The probable explanation is that, while parsing research focused on invented examples, these involved mainly core language constructions whose analysis was studied by theoretical linguists, so it was not felt necessary for computational linguists to discuss the issue independently.

However, real-life language samples include many phenomena (one example would be postal addresses) whose structure is of little interest to linguistic theory. Furthermore, theoreticians frequently disagree about the analysis of core constructions, whereas treebanks require some one consistent analysis to be imposed – arbitrarily if necessary. In the 1990s, attention began to be paid to this area, with publication of rigorously-defined and comprehensive parsing schemes, notably those of the Penn and SUSANNE treebanks (Bies et al., 1995; Sampson, 1995).

3. Parse-accuracy metrics

In the early years of automatic parsing, using mainly invented and unnaturally simple language samples and fairly austere schemes of analytic categories, it may have been reasonable to assess parser output in terms of percentage of input sentences correctly parsed, without quantifying degrees of inaccuracy in the case of parses containing errors. Once parsers deal with realistically complex input samples and use the refined parsing schemes that are now usual, on the other hand, simply classifying sentence parses as "right" or "wrong" becomes clearly inadequate. A parse which differs from the gold standard for the relevant input only in terms of a discrepancy in one attribute of some low-level grammatical tagma is a very different matter from a parse whose tree structure is quite unlike that of the gold standard, and where a node that coincides (in terms of word-sequence dominated) with some gold-standard parse node is often labelled with a fundamentally different grammatical category. Both are incorrect parses, but from either a theoretical or a practical point of view the former incorrect parse is much closer to the gold standard than it is to the latter incorrect parse. We need metrics for quantifying degrees of parsing accuracy, which should preferably not just assign global figures of merit to sentence parses but should also include detailed information about the location and nature of parsing errors.

For some years the task of developing satisfactory metrics for parsing accuracy was held back by the accidental fact that a particular metric that is now generally recognized to be unsatisfactory became accepted as a de facto standard by the research community. The PARSEVAL parser-evaluation competition series used a metric, called either the GEIG – "Grammar Evaluation Interest Group" – or simply the PARSEVAL metric (Black et al., 1991; and further developed in Magerman, 1995, Collins, 1997), whose nature was heavily influenced by the need for it to enable outputs from different research groups' parsers, based on very diverse parsing schemes, to be compared on equal terms. The use of this metric in an international competition series gave it standard status, even in contexts where comparing parsers using different target parsing schemes was not relevant. But the numbers yielded by this metric do not correlate well with people's intuitions about relative goodness or badness of parses – as Bangalore et al. (1998) put it, "it is unclear as to how the score on [the GEIG] metric relates to success in parsing". Two workshops (Carroll, 1998; Carroll, 2002) have recently been devoted to criticizing and moving beyond the GEIG metric, and alternatives to it (e.g. the "leaf-ancestor" metric, Sampson, 2000) have been developed.

4. Human parsing performance

Even if we have well-defined and comprehensive parsing schemes, automatic parsers that perform well in identifying how such schemes apply to particular real-life language samples, and refined parse evaluation metrics which yield meaningful and detailed information about the imperfections of parser performance, there is still a fourth area which is logically needed for a complete picture – but to which very little attention has been paid to date. Namely: how precisely can human beings analyse language structure? An automatic parser is a mechanical system for making explicit the structure which human language users implicitly assign to a language sample in producing or understanding it. One can hold philosophical debates about whether language means what speakers and writers intend it to mean, or what hearers and readers take it to mean (though differences there are likely to be too subtle to have much relevance for the current state of computational linguistics); but no-one (surely) would dispute that human performance is the ultimate criterion for automatic language analysis.

To draw an analogy with another area of computational linguistics, namely machine translation, it would not make sense to claim that some MT system was capable of translating language A into language B *better* than the best human translators for that language-pair: skilled human performance logically defines an upper bound for machine performance. Different human translators, or the same translator on different occasions, will often produce non-identical target-language renderings of a given source-language text, but if the people in question are fully competent this simply shows (what we all know is true) that there is an unavoidable looseness about the process of translation. Some MT system might produce a translation which matched one human translation at some points and the other human translation at other points, but we could not think of it as "adjudicating" between the two translations by showing where the respective human translators had "gone wrong". What justification would there be for preferring the machine's rendering of a phrase to the output of whichever skilled human translator chose a different rendering at that point?

This kind of consideration may have seemed irrelevant for grammatical analysis during the pre-treebank period of parser research. While automatic parsers focused on detecting core linguistic constructions in artificially simple invented examples, there was little reason to think about the limits of human parsing performance. Indeed, in one respect it did make good sense at that period to talk about automatic parsers as performing better than humans: automatic parsers will commonly list all possible analyses of a structurally ambiguous sentence, whereas a human reader will normally notice only one or a small number of plausible interpretations, while being willing to agree that many other interpretations are possible if they are drawn to his attention. One well-known example is in Martin et al. (1987), who point out that their grammar assigns three analyses to the sentence *List the sales of products in 1973.*, but (because of combinatorial explosion) assigns 455 alternative analyses to *List the sales of products produced in 1973 with the products produced in 1972.* No human hearing or reading the latter example is likely spontaneously to think of a fraction of that range of interpretations.

Ambiguity is logically a distinct issue, although one that is hard to keep separate from the issue under discussion here. Once language-samples, and the parsing schemes developed to annotate them, display the complexity and refinement that become appropriate with treebanks representing real-life usage, then real issues arise about how much detail can meaningfully be assigned by a human annotator even to an unambiguous sample, or to an ambiguous sample in any one of its possible interpretations.

For instance, the SUSANNE scheme provides three alternative analyses for cases where a word having the form of a past participle, e.g. *involved*, follows part of the verb *BE* (Sampson, 1995: 130–1, 262). The word may be treated as an adjective (e.g. in *the wording was very involved*), as the head of a nonfinite clause (e.g. in *they were involved in various shady dealings*), or *BE* together with the participle may be treated as a passive construction (e.g. in *we are being involved despite our protests*). The (invented) examples quoted here contain cues which "force" one analysis in each case, but it is easy to imagine that these distinctions may be too subtle to be drawn consistently in many real-life cases – meaning not that real-life examples are ambiguous, but that requiring analysts to choose one of these three analyses rather than the other two would often be asking them to answer non-questions.

For scientific adequacy we want our annotation scheme to show all the structural detail that is really there in the language, but we do not want to adopt annotation schemes that ask analysts to draw distinctions which in reality trained human language users cannot make reliably. We need to establish an upper bound on parse-precision. Information about the level of detail in the annotations used in a particular treebank, or about the degree of accuracy of a particular automatic parser, is only really enlightening in the context of benchmarks permitting comparison with skilled human performance.

5. Existing work

Some work of this kind has occurred, but very little.

In the first place, definitions of parsing schemes sometimes make explicit that decisions to recognize one logical distinction but not another are based on experience of which distinctions analysts are capable of drawing reliably. For instance, the annotation scheme of the Switchboard Corpus of telephone conversations (Meteer et al., 1995) contains a number of remarks such as "annotators were basically unable to distinguish the discourse marker from the conjunctive use of *so*". Comments like these shed sporadic rays of light on the limits of human performance in particular analytic areas, but they do not amount to (and are not intended to amount to) a systematic survey of the limits to human analytic precision across the board of language structure.

In their series of experiments with alternative techniques for automatic resolution of PP-attachment ambiguities, Hindle and Rooth (1993) are careful to establish a baseline of human performance against which to evaluate results from the different mechanical techniques. Since the "incorrect" attachments are commonly readings which in principle would be linguistically possible although not plausible or not intended by the writers, this research has more to do with choice between alternative legal interpretations of structurally-ambiguous wording than with the question of which analytic categories can reliably be recognized as applying to wording in a single interpretation. Nevertheless, it is an interesting precedent for the kind of upper-bound identification which is too often missing when researchers discuss treebank annotation or automatic parsing of real-life language samples.

Babarczy et al. (2001) report an experiment designed to establish an upper bound to inter-annotator agreement for the task of wordtagging samples from the written section of the British National Corpus, using a highly-refined tagset. Our figures distinguish between various causes of non-agreement: ambiguity, where analysts had chosen different tags each of which was correct with respect to one of alternative interpretations of the text, is differentiated from inability of analysts to agree on the same tag for a single interpretation (and the latter cases are in turn broken down under separate subheadings). Wordtagging is of course only one small aspect of the total task of structural annotation of language samples, but this experiment illustrates the type of investigation that is needed.

Doubtless the literature contains further relevant material overlooked by the present writers. Nevertheless it seems true to say that the issue of establishing the upper bound to parser performance in terms of trained human ability is to date a severely neglected issue, as compared with the three other issues of defining parsing schemes, designing automatic systems for parsing in accordance with those schemes, and measuring the performance of the automatic systems (the best of which at present surely fall well short of the upper bound).

This neglect may seem as ill-advised as it would be to design systems for automatically generating original musical compositions, without also studying in detail what kinds of arrangements of sounds appeal to human musical sensibilities. There are other cases where it makes good sense to design machines to do a job without considering how far humans or other natural organisms are capable of doing the job. Nobody expects aeronautical engineers to think about birds when designing aeroplanes; that is because the task achieved by an aeroplane is well-defined independently of the fact that creatures exist in Nature which carry it out to some extent. On the other hand, an arrangement of sounds is musical only if human beings hear it as musical; there is no independent definition of musicality. Structural analysis of language is in this respect like musical composition, not like flying.

6. A planned investigation

The chief purpose of the present paper is to urge the research community at large to take seriously the task discussed, and to encourage others to engage with it. Nevertheless, it may be worth adding a brief description of a pilot experiment which the present writers are about to undertake in order to get an initial handle on the issue.

The experiment will involve comparing the output of two human analysts (the present authors) applying the same parsing scheme independently to the same language samples. For such an experiment the parsing scheme needs to be as comprehensive and tightly defined as possible (the results will not be very instructive unless the human analysts prove incapable of matching one another's output in some respects). We shall be using the SUSANNE scheme, which a number of commentators have described as more refined than any other (e.g. Terence Langendoen, 1997: 600: "the detail … is unrivalled").

Clearly, in an ideal world it would be preferable to use more than two analysts. But it is essential to minimize the likelihood of discrepancies arising merely because one or both analysts are not as well-versed in the scheme they are trying to apply as they might be, and in practice this means that having two suitable researchers available at the same time and place, both with long experience of working with the same detailed annotation scheme, is about as much as can reasonably be hoped for. Note that the experiment is not about the speed or accuracy with which newcomers can learn to apply a detailed natural-language annotation scheme (those might be interesting topics, but are not under investigation here); it is about how predictable annotation of real-life language can be, given a maximally detailed scheme of guidelines and annotators who are as familiar with that scheme as can reasonably be hoped. The intention is to establish a ceiling on human annotation accuracy, which in less favourable circumstances will not always be attained.

We shall use samples of edited written language, drawn at random points from appropriate sections (e.g. published rather than informal, unpublished material) of the British National Corpus. In principle, we would like to explore human performance on all genres of language, including spontaneous speech and informal, unedited writing; but, for a pilot experiment with limited resources, it seems best to use a genre where the problems of analysis relate to definition of the analytic categories, rather than to clumsiness of speaker performance.

We currently envisage using a total of 20,000 words drawn in two-thousand-word chunks from a variety of BNC published texts. This quantity is chosen to be large enough to constitute a valid test of analytic reliability, while small enough to permit detailed examination of individual discrepancies.

(Note that, in a field like this where at present we have little idea what results we shall find for various aspects of structure at even an order-of-magnitude level, it will be more useful to analyse the findings thoroughly in terms of nature and source of different discrepancies than to process sufficient quantities of material to add an extra significant decimal place to the numerical results.)

The BNC, although the most suitable data-source we have, is flawed in many respects – e.g. imperfect OCR output uncorrected, printed representations of dialogue "normalized" with insufficient appreciation of the diversity of publishers' typographic conventions. Divergencies between annotators in reacting to these kinds of problem are of no theoretical interest and would only blur the focus on discrepancies in treatment of valid input. Therefore the experiment requires initial work not only to recast the BNC selections into the format required by our existing annotation-support software tools, but also to get them into a state where remaining oddities in the spoken and written texts reflect oddities in the original language, rather than errors in the process of compiling the BNC.

7. Implementation of leaf-ancestor assessment

The independent annotated outputs of the two analysts will be compared using the leaf-ancestor metric; an experiment reported by Sampson and Babarczy (2002) suggests that this succeeds well, and certainly much better than the GEIG metric, at measuring analytic discrepancies in a way that accords with intuitive judgements of their relative gravity (that is, it gives low marks to "bad mistakes" and high but not perfect marks to "minor errors"). The usual application of a parse-accuracy metric is to compare the output of an automatic parser with a hand-crafted "gold standard" parse; in the present experiment, the metric will be applied in a more symmetrical manner, with neither analyst's output treated as more authoritative than the other's, but this involves no change to the operation of the evaluation software itself.

This software will be based on the program written for the Sampson and Babarczy (2002) experiment, which returns quantitative measures of similarity between alternative labelled trees over the same language sample, computed as described in Sampson (2000). That program will be refined, extended to yield a fuller range of information, and documented. The extended software will include user-controllable parameters allowing particular aspects of node-labels to be ignored in comparing label-pairs, or labels of particular categories to be differentially weighted.

Since one of the relevant analytic issues is division of written text into taggable words (what computer scientists often call "tokenization", though properly speaking this is a misunderstanding of Peirce's type/token distinction) in the case of hyphenations, unspaced sequences like *£25*, *10%*, etc., the software will also need to be able to compare parse-trees in which leaf-nodes correspond to individual characters and words occupy nonterminal nodes, though the option of comparing orthodox trees with words at the leaf-nodes will also be available in circumstances where "tokenization" is not at issue.

It is hoped that a by-product of this experiment will be software implementing the leaf-ancestor parse assessment metric engineered to a standard suitable to be placed in the public domain.

8. Measurement of inter-annotator agreement

The leaf-ancestor assessment software will be used to measure annotator agreement on the spoken and written texts, globally and with respect to various features of structural annotation, such as form- v. function-tagging, analysis of speech repairs v. fluent passages of speech, high-level (near root nodes) v. low-level (near leaf nodes) structure, "core" grammar v. structure in e.g. addresses or money sums. Initial findings about areas of discrepancy will guide the experimenters to more specific parameter settings; the outcome will be a series of numerical measures showing how inter-annotator agreement varies with language genre, aspect of language structure, and type of construction.

A particular advantage of the leaf-ancestor metric is that as well as giving global figures for the similarity of different parse-trees over the same string, it also yields information about local accuracy of parsing on a word-by-word basis. This will facilitate a statistical investigation of the nature of specific inter-analyst discrepancies, in order to discover how far these correspond to alternative interpretations of genuinely ambiguous structure and how far they reveal that the formalisms of the annotation scheme have outrun the ability of skilled human analysts to apply them consistently to represent one particular interpretation.

Manual analyses of inter-annotator discrepancies will be conducted, classifying individual discrepancies as:

(1) the language is inherently unclear/ambiguous

(2) the language is clear but the guidelines are vague/missing/contradictory; it would be possible to extend the guidelines to give a predictable analysis in such cases

(3) as (2), but it would be difficult to devise a suitable extension to the guidelines to handle such cases

(4) the language and the guidelines are unambiguous, but one or both annotators failed to apply the guidelines correctly

Even cases of type (4) are worth counting; the complexity of natural-language structure is such that one important factor determining the ceiling on annotation accuracy is the extent to which even experts can hold a comprehensive scheme of guidelines in their head. But this factor needs to be differentiated from the factor of inherent vagueness in language structure. It may emerge that some areas of the SUSANNE annotation scheme already make finer distinctions than it is possible to apply consistently in practice, while in other areas of structure the existing categories may prove to be well-defined and greater analytic refinement would be possible. Cases of type (2) will lead to published improvements in the existing annotation scheme.

9. Conclusion

For purposes of specific language-processing applications, various well-defined aspects of structure may be irrelevant, but system developers need to know how far it makes sense to be precise about whichever aspects of language structure are relevant to a given application. The general enterprise of natural-language parser development presupposes understanding of how far human language-users are capable of detecting grammatical distinctions.

Work designed to develop such understanding, as exemplified by the pilot study discussed above, is by now overdue.

References

Babarczy, A., et al. 2001. Annotator error rates for part-of-speech tagging. LINC 2001, at 34th SLE, Leuven, Belgium, 28 Aug–1 Sep 2001.

Bangalore, S., et al. 1998. Grammar and parser evaluation in the XTAG project. In Carroll (1998).

Bies, A., et al. 1995. Bracketing guidelines for Treebank II Style, Penn Treebank Project. www.cis.upenn.edu/~treebank/home.html

Black, E., et al. 1991. A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the Speech and Natural Language Workshop, DARPA, Feb. 1991, Pacific Grove, Calif.*, pp. 306–11. Morgan Kaufmann.

Carroll, J.A., ed. 1998. *Proceedings of Workshop on the Evaluation of Parsing Systems*, at 1st LREC, Granada, Spain, 26 May 1998 (Cognitive Science Research Papers 489, University of Sussex).

Carroll, J.A., ed. 2002. *Proceedings of Workshop "Beyond Parseval – towards improved evaluation measures for parsing systems"*, at 3rd LREC, Las Palmas, Canary Islands, 2 June 2002.

Charniak, E. 1996. Tree-bank grammars. In *Proceedings of the 13th AAAI*, Portland, Oregon, vol. 2, pp. 1031–6.

Collins, M. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th ACL and 8th EACL, 7–12 July 1997, Madrid*. Morgan Kaufmann.

Hindle, D., and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics* 19.103–20.

Langendoen, D.T. 1997. Review of Sampson 1995. *Language* 73.600–3.

Magerman, D.M. 1995. Statistical decision tree models for parsing. In *Proceedings of the 33rd ACL, 28–30 June 1995, Cambridge, Mass.*, pp. 276–83. Morgan Kaufmann.

Martin, W., et al. 1987. Preliminary analysis of a breadth-first parsing algorithm: theoretical and experimental results. In L. Bolc (ed.), *Natural Language Parsing Systems*, Springer (Berlin).

Meteer, M., et al. 1995. Dysfluency annotation stylebook for the Switchboard Corpus. `www.ldc.upenn.edu/Catalog/CatalogList/LDC99T42/DFLGUIDE.PS`

Obermeier, K.K. 1989. *Natural Language Processing Technologies in Artificial Intelligence: the Science and Industry Perspective*. Ellis Horwood (Chichester, Sussex).

Sampson, G.R. 1995. *English for the Computer: the SUSANNE Corpus and Analytic Scheme*. Clarendon (Oxford University Press).

Sampson, G.R. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics* 5.53–68. Online at `www.grsampson. net/Apfi.html`

Sampson, G.R. and A. Babarczy. 2002. A test of the leaf-ancestor metric for parse accuracy. In Carroll (2002); to appear in *Natural Language Engineering*. Online at `www. grsampson.net/Atot.html`

Spärck Jones, K., and Y.A. Wilks. 1983. *Automatic Natural Language Parsing*. Ellis Horwood (Chichester, Sussex).