Doctoral thesis report

Study programme: Informatics
Field of study: Informatics
Name of student: Vojtěk Kovář
Thesis title: Automatic Syntactic Analysis for Real-World Applications

Reviewer: John Carroll
Institution: University of Sussex

The thesis is concerned with the automatic syntactic analysis of text, investigating appropriate methodology and techniques, motivating these from a language-based application perspective and assessing their utility also from this angle. It presents a novel approach to developing parsers for language-based application tasks, in which the practical requirements of the task take precedence over optimisation of accuracy in intrinsic evaluations of parser output against gold standard syntactic analyses. The approach is demonstrated in several practical scenarios. A good amount of high quality technical work has been done; this work has led to a high quality collection of publications. The thesis makes a significant contribution to learning in the field of natural language parsing and language-based application development.

The thesis is organised well. The final chapter summarises the main findings, but should also propose directions for future work.

The thesis starts (Chapter 1) with a statement of the goals of the research and an outline of various kinds of language processing applications that could benefit from automatic syntactic analysis.

Chapter 2 discusses fundamental practical issues in syntactic analysis: the preprocessing steps necessary before raw text can be input to a parser, ways in which syntactic information can be represented, computational techniques used for parsing to produce such information, widely used tools that implement these techniques, and methods for evaluating parsers. The discussion about syntactic representations focusses on the two main approaches, phrase structure and dependencies, pointing out problems with both, variously stemming from intrinsic properties of the representations as well as choices in how the representations are used to capture specific linguistic phenomena. The thesis mentions but does not discuss in detail syntactic theories and associated tools producing more detailed information which could support principled semantic interpretation (e.g. LFG, HPSG). Some of these tools have been used in real-world, commercial applications – in areas including machine translation, document routing and question answering; they are relevant to the topic of the thesis and deserve attention. The chapter goes on to argue why parsers appear not to be used by the NLP research community in support of other language processing tasks. Evidence for this is presented based on a regular expression search in research papers for a specific phrase indicating that a parser has been used; however, this search will produce many false negatives. Although the argument is sound for some kinds of application, the overall picture is uneven. As a random example, 6 out of the 14 teams who entered

the Native Language Identification Shared Task at the 8th Workshop on Innovative Use of NLP for Building Educational Applications in 2013 used features derived from syntactic analysis. The chapter concludes with a good discussion of syntactic analysis accuracy and task accuracy; a further relevant point is that some dependency-based evaluation metrics can distinguish between different kinds of parser error, providing a finer grained way to investigate the relationship between parser error and task error.

Chapter 3 describes an approach to building parsers that are targeted to application tasks and domains, taking as a starting point the standard syntactic annotation of text (as for example in the Penn Treebank and Prague Dependency Treebank) but with a highly simplified representation using both phrases and dependencies. This 'syntactic bush' representation is shown to be much cheaper for annotators to produce, since they require less training and can produce the required annotations faster. Although it is the case that standard, widely used treebanks contain edited text in a single genre from 20 years ago, there are recent approaches which attempt to overcome the same problem as tackled here; these include the SANCL initiative, domain adaptation techniques, and augmentation of an out-of-domain treebank-derived parsing model with small amounts of in-domain treebanked text. An interesting, novel feature of the new approach is a framework for annotation of the same text by multiple annotators, which allows training and evaluation to take into account annotation confidence levels; this approach is related to Carroll et al.'s investigations into weighted grammatical relations (COLING 2002, IWPT 2005), but generalises it from parsing to syntactic annotation and statistical parser training. The chapter finally presents an initial evaluation of parsers against the syntactic bush data, showing that good accuracy can be achieved.

The focus of Chapter 4 is a different, but still application-driven approach to parser development. The parser forms part of the Sketch Engine application, whose primary users are lexicographers. The parser is built around a 'sketch grammar', which is a set of patterns and processing directives written in a corpus query language. Extensions to process multiwords and produce bilingual word relations are described. The parser is used by thousands of people – although it is debatable whether this breadth of parser usage is as unique as is claimed. (The system for automatically marking English language learner essays at http://www.flo-joe.co.uk/fce/students/writing/assessment.htm incorporates a parser, and is heavily used worldwide).

Chapter 5 is about a further, more generic parsing system, SET. The system was developed according to the hypothesis that good evaluation results against a gold standard are less important than simplicity in design, adaptability to different applications, and comprehensibility of output. The parser output combines phrase and dependency structure, as in the bush representation of Chapter 3; however this 'hybrid tree' output is formalised differently without an explanation how it relates to the earlier representation. The SET parser uses a collection of template-action patterns in a similar way to the sketch grammar parser. The depth of recursion is limited; this is justified with reference to Miller's psychological investigations, although Sampson's depth in English grammar findings (*Journal of Linguistics* 1997) are more relevant. Pattern-based partial parsing has been successfully employed before, most notably by Abney, Grefenstette and the Helsinki Constraint Grammar team led by Karlsson in the 1990s; a comparison with these approaches would be informative. Although the parser's speed of 7 sentences per second will be sufficient for some applications, it is insufficient for some kinds of streaming data, e.g. real-time Twitter feeds on globally trending topics. Other current parsers – such

as the C&C system – are faster by up to an order of magnitude.

Chapter 6 applies the sketch grammars and SET parser of Chapters 4 and 5 in a number of real-world language processing applications. This provides a practical trial of the claim in Chapter 3 that the flexibility, adjustability and simplicity of these parsers makes them more suitable for applications than statistical parsers trained on treebanks. In each case the parsers are customised to better process specific text types or meet particular application requirements. Successful use of the parsers demonstrates that the there is good basis to the claim. However, the evidence would be stronger if there were side-by-side comparisons with adapted versions of other parsers, with the same amount of effort being invested in these adaptations as in producing customised versions of SET.

Chapter 7 gives a short summary of the main findings. Overall, the thesis raises and provides some interesting answers to important questions about effective development of real-world language processing applications and how parsing can contribute to them.

The thesis expresses ideas clearly. It contains discussions of relevant literature on parsing and natural language processing applications. However, in a few cases (as indicated above) relevant previous work is overlooked. The bibliography is accurate. The thesis is well-presented with good use of diagrams and figures.

In the defense, the candidate could be asked to comment on that fact that the thesis comes to conclusions that differ from analogous cases in building other kinds of computer systems; software engineers look for abstractions of the requirements of classes of applications so that they can build reusable components. (For example, when implementing a graphical interface for an application we don't write the graphics code from the ground up, but use an existing graphics library with well-defined interfaces). In a similar way, should language processing application developers be encouraged to identify requirements for syntactic information that are common across classes of applications?

The author in his dissertation thesis has demonstrated the ability to work independently and creatively in the specified field.

The thesis meets the standard requirements imposed on the dissertation thesis in the field.

I recommend that the thesis be accepted.

I grade the thesis as B.

Masarykova univerzita
Fakulta informatiky
Datum: 11 -08- 2014
Č.j./E.č.: .....
Počet listů dokumentu: .....
Počet příloh a listů/sv.: ...../.....
Počet a druh nelist. příloh: ...../.....

John Carroll
7th August 2014