REPORT ON DOCTORAL THESIS

Title: Unsupervised Dependency Parsing

Author: David Marecek

Opponent: Ing. Filip Jurcicek, Ph.D.
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranske namesti 25
118 00 Prague 1
Czech Republic

In general, the aim of dependency parsing is to identify relations between words in a sentence. There are two typical approaches to this problem. The first approach is to manually construct as set of rules by an expert, e.g. a linguist, which are then applied during parsing. The second approach is to build a statistical dependency parser using some machine learning technique which learns its model from a corpus of previously annotated data. Both approaches rely on linguistic knowledge either in the form of rules or an annotated corpus. These techniques are called supervised since experts have to create either the rules or a corpus.

The topic of the submitted thesis is unsupervised dependency parsing. In contrast to the previously mentioned techniques, the thesis presents machine learning techniques used to build a statistical model without supervision (a corpus of annotated data) and automatically parse input sentences and assign dependencies between the words. In other words, the author proposes a method which could alleviate the burden of manual creation of training corpora. **In this sense, this topic is up-to-date and it is very desirable to find a solution to this problem**.

The thesis starts with an introduction of the topic, the forms of representing the relations between the words, and the goal of the thesis. Then, the next chapter describes the related work. The third chapter describes the basics of employed machine learning techniques – namely Bayesian inference and its practical implementation in the form of Gibbs sampling. The fourth chapter describes the available training data and evaluation metrics. The next chapter presents the proposed model of the unsupervised parser. The sixth chapter describes the inference using the previously proposed model. The extensive evaluation on more than 20 languages is presented in the seventh chapter. The last chapter concludes the thesis and summarize the main results.

The thesis is written in concise language and it is easy to read and understand. However, there are certain points which might be improved:

1) The author uses the terms projective and non-projective trees. Although it is rigorously defined on page 6. It would be useful to have some examples of projective and non-projective trees in the same section.

2) The term “directed attachment score” (scores in general) is already used on the page 14, however a reader must wait for its formal definition until the page 31. It would help to describe the score earlier.

3) On several occasions more details should be provided to allow accurate re-implantation by a reader, e.g. the section 6.2.2.

In addition, I have several more comments:

1) Section 3.3: the author comments on the effect of the choice of the α parameter of the symmetric Dirichlet distribution. The author writes “With increasing α, parameters θ approaches towards having the uniform distribution.” However, this is not very precise interpretation. When assessing θ, we should talk about expectations of θ since θ is a random variable. The correct understanding of θ is: the expectation of θ (E(θ)) always forms a uniform distribution and the variance of θ converges to 0 with increasing α. Therefore for $\alpha < 1$, the Dirichlet prior is weak (allows for high variance of θ) and it can converge easily to different solutions including a sparse solution. For $\alpha >> 1$, the prior dominates the data and therefore the posterior will not have a sparse solution. Can the author comment on this?

2) Section 3.3.1: the author relates the Dirichlet distribution to the Chinese restaurant process. However, the Chinese restaurant process is an infinite process and therefore it is equivalent to the Dirichlet process which is an extension of Dirichlet distribution for infinite spaces. Can the author comment on this?

3) Section 3.4: the author briefly explains the principles of the Gibbs sampling method used in Bayesian inference. This section would deserve some expansion since the Gibbs sampling is used as the main machine learning technique in this work. Also the author should introduce the technique called collapsed Gibbs sampling. The technique which is the author really using – integrating out the model parameters and sampling only dependency relations.

4) Section 5.1.3: the author operates with the history; however, it is not completely clear what the history exactly is at this point. Is it all the data in the corpus, is it all the data without *d*, or is it only the data appearing before d (assuming that all the word in a corpus is ordered left to right)? Can the author comment on this?

5) Section 5.1 and 5.2: in my opinion, only the edge and fertility models are truly unsupervised statistical models. The distance and reducibility models are in fact heuristic models designed by an expert linguist especially because their parameters ϒ and δ are tuned on an English treebank using a grid search method (as described in Section 7.3.2). Nevertheless, their nice property of the distance and reducibility models is that they work across many languages. So, the author can claim that they are at least language independent. Can the author comment on this?

6) Section 5.4.2: since the model of reducibility is a one of the main results of the thesis then it would be helpful to include some examples of how the scores are computed in practise. The definition of the computation is precise; however it is not trivial to imagine especially because of combining counts and testing on POS tag and word sequences.

7) Section 6.1: the author defines probability of a treebank (eq. 6.1). Later, it is used to sample from the dependencies in the inference algorithm 6.3 by computing the probability of the Treebank for differently placed dependencies in the corpus and then normalising it to obtain a probability distribution. However strictly speaking, in Gibbs sampling, one should derive a posterior probability for a dependency relation e.g. an edge being or not being between two words and then sample from it. Can the author comment on this?

   This also relates to the example in the equation 6.2. Here it appears that the author use really only the "left part" of the corpus when computing likelihood of a specific dependency. However, in Bayesian inference we can use all available data including the "right part" part of the corpus since it is available in at the time of computation. Can the author comment on this?

   However as seen from the eq. 6.5, it seams that both left and right parts of a corpus are used by denoting *others* in the counts.

   The author also refers to exchangeability which is in some sense related to the dependencies being independent and identically-distributed (i.i.d) random variables (however not the same). Although the dependencies are very likely to be i.i.d between different sentences, they are not for sure i.i.d within one sentence. In one sentence, adding one dependency forces other dependences to change to maintain a tree structure in the treebank. Did the author take this in to consideration?

8) Section 6.2.2: the section does not describe how exactly sample the changes. It does not contain derived distributions for the small change operator. One cannot re-implement the proposed technique – especially because of the bullet point 3. Can the author comment on this?

The amount of cited literature is acceptable and it is apparent that the author is aware of all the basic and the state-of-the-art methods. The proposed technique was evaluated extensively and **it is clear that the proposed method delivers excellent results**. The scientific progress can be seen in the following areas: 1) definition of the reducibility model, 2) development of a small change operator.

Evaluation: In my opinion, the author showed in this thesis that he is capable of independent research and **the presented results have broad impact**. Therefore **I recommend this thesis for a defence**.

In Prague, 3. 8. 2012.

Filip Jurcicek