Probabilistic Grammars and Data-Oriented Parsing
Spring 2006
Short Essay Submision Guidelines

Reut Tsarfaty

Institute for Logic Language and Computation
University of Amsterdam

1 Procedure

1.1 General

- As part of the completion requirements for the course, every student has to submit a weekly short essay about the preceding lecture, following up on a question posed by the lecturer on the course website.
- The essay should be phrased in your own words.
- The essay should not exceed 300 words, excluding mathematical formulas and references (when applicable).
- The essays should be sent to `rtsarfat@science.uva.nl` before the next lecture, or handed in at the beginning of the lecture, to me or to one of the lecturers. If you foresee a problem with meeting this deadline please contact me in advance. (Late submissions which were not coordinated with me in advance may be penalized.)
- Submission of an essay is possible only for students who attended the lecture.

1.2 Grading

- In order to pass the course at least 10 essays should be submitted.
- The grade for the essay submission will constitute 30% of your final grade.
- The individual essays will be graded according to the following scale

A+	10	B+	8.5	C+	7	D	failed
A	9.5	B	8	C	6.5	-	-
A-	9	B-	7.5	C-	6	-	-

- Your 10 best grades will be used to calculate the semester's average.

2 Evaluation

2.1 Criteria

The individual papers will be evaluated based on the following criteria:

- Content
  - Does the essay answer the question?
  - Does the essay provide sufficient explanation for related technical terms?
  - Does the essay demonstrate sufficient level of understanding of the subject matter?
- Correctness
  - Correctness of definitions.
  - Correctness of mathematical formulas, pseudo-code and algorithms.
  - Correct and appropriate use of terminology.
- Writing
  - Length requirement
  - Writing style
  - English

3 How to write?

Before you write, ask yourself the following questions:

- What is the main problem we addressed in class? (in what general context did it appear?)
- What is the general approach towards the solution? (does it contrast with other existing ones?)
- What is the essence, the key idea, of the solution?

- What are the relevant technical terms?
- What are the substantial mathematical terms, formulas, or algorithms?

When you write, make sure that:

- All of the above items are covered sufficiently.
- All items are integrated into a flowing narrative that reflects the structure of your argument.
- Specify exact formulas/algorithms, however not on account of your own explanation.
- Be precise, don't use vague or general phrases.
- Write appropriately (check spelling, grammar).

4 Example: HW set #1

Read the question carefully:

> What is a language model over parse trees? how do you estimate the PCFG probabilities from a treebank?

Mark important keywords:

> What is a **language model** over **parse trees**? how do you **estimate** the **PCFG probabilities** from a **treebank**?

Think of relevant questions you could ask yourself, and make sure you know how to answer them. Here is an example for a list of questions that could be compiled based on the the above keywords.

- What are probabilistic language models?
- What are probabilistic language models over structures?
- What are probabilistic language models over parse trees?
- What is a parse tree?
- What is a PCFG?
- What is a treebank?
- What is the goal of estimation?
- What are we trying to estimate?
- What kind of resources do we need for estimation?

- What functions/values need to be defined/calculated for the purpose of estimation?

Write a skeleton that puts together all the pieces:

> Probabilistic Language Models over linguistic structures are ....
> In particular, we can define probabilistic language models over syntactic structures, typically referred to as 'parse trees', which are .....
> Syntactic structures can be expressed using a Context Free Grammar (CFG), which defines ..... .
> A Probabilistic Context Free Grammar (PCFG) augment the production rules of the grammar with ....
> The probability of an utterance can be calculated using a PCFG ...
> the probability of a parse tree is ....
> the probability of a derivation is ....
>
> Since probabilities are theoretical notions, we need to estimate the parameters ....
> Assuming a language sample was generated by an underlying PCFG we can use a treebank for ...
> A treebank is ....
> When complete data is available (the 'supervised' case) we can estimate the probabilities of the PCFG by calculating the Relative Frequency Estimate (RFE) of the production rules, which is .....
> When the data is incomplete (the 'unsupervised' case) , we can apply EM methods, which are .....

Finally, fill in the missing details, spell check, proof read, correct, count the words and submit. Here is an example for an essay for homework 1.

> Probabilistic language models over linguistic structures are probability distributions over utterance-analysis pairs. Specifically, a probabilistic language model over syntactic structures is a probability distribution over utterance-parse tree pairs, where parse trees represent the syntactic structure of sentences. One way to define syntactic structures is by using a Context Free Grammar (CFG) that defines a set of the production rules, in which the left hand side (LHS) can be substituted for the right hand side (RHS) independently of context. A Probabilistic Context Free Grammar (PCFG) augment the production rules of a PCFG with probabilities, allowing to calculate the probability of a parse tree by multiplying the probabilities of all rules used in a derivation. The probability of an utterance is then defined as the sum of the probabilities of all possible derivations that yield a given utterance. (In PCFGs, as opposed to e.g. STSGs, the probability of a derivation equals the probability of a parse tree).
>
> Probabilities in probabilistic grammars are abstract mathematical ob-

jects (parameters), whose values are yet to be estimated based on real data. Assuming a language sample was generated by an underlying PCFG, we can use a treebank, i.e., a set of sentences annotated with their corresponding syntactic structures, to learn the underlying PCFG. We can extract the CFG rules used in the derivations, and estimate their probabilities using a 'treebank training'. When complete data is available (the 'supervised' case), we estimate the probabilities by calculating the Relative Frequency Estimate (RFE) of the production rules. This is achieved by dividing the occurrence frequency of each rule by the sum of occurrence frequencies of all the rules that share the same RHS, i.e., $\frac{C(A \rightarrow \alpha)}{\Sigma_{\{\beta | A \rightarrow \beta\}} C(A \rightarrow \beta)}$. When the data is incomplete (the 'unsupervised' case), we can apply Expectation Maximization (EM) methods, which will be treated separately in subsequent lectures.

If you have further questions do not hesitate to contact me

rtsarfat@science.uva.nl

Good luck !!!