Statistical Natural Language Processing (SS-2018)
Exercise Sheet 9

Submission Deadline: 22.06.2018, 23:59 (CEST)

Markov Random Fields (MRFs)

1) (2 points) Let X, Y, Z be three disjoint subsets of random variables (RV). Suppose

$$p(x, y, z) = h(x, y)g(y, z)$$

- Draw the corresponding MRF
- Does it imply $p(x|z) = p(x|z, y)$?
- If yes, prove it. Otherwise, please provide a counterexample.

Conditional Random Fields (CRFs)

1) (1+1 points) List 5 features that could be useful for the task of Named Entity Recognition (NER). Also, justify how the above listed features could be useful for improving the accuracy of NER.

2) (1 point) List and briefly explain at least two differences between CRFs and Hidden Markov Models (HMMs).

NER using CRFs

In this task, you will build an NER system using CRFs. For this, we will make use of the software *CRF++* toolkit. First download and install the software by following the instructions specified at the CRFPP webpage.

And, before diving into the actual task, you might want to read about it here: CoNLL-2003 Language-Independent NER. Now, use the given data '*data-eng/**' for training and developing the model. The data is already in a proper format. Read about how to use this data as input to CRF++ here: Usage Instructions.

1) (0.5 points) Use as many features as you want.

2) (0.5 points) Use as many feature functions as you want.

3) (3 points) Implement & train your model.

4) (1 point) Report *Precision, Recall*, and *F1-Score* on the dev data and comment briefly about them.

We provide you only the train and dev data. We reserve the test data for testing your model. Submit the code as usual by e-mail to respective tutors. Due to size limitations, upload your final *model* file to some cloud drives and share the link in the report.

We will test your model (on the reserved test data). Among all the teams, top 3 teams would be rewarded with **bonus points** as follows:

1^{st} position : 7 points
2^{nd} position : 5 points
3^{rd} position : 3 points

For the inquisitive minds, more information about the task can be found from the original paper, Introduction to the CoNLL 2003 Shared Task: Language-Independent Named Entity Recognition

---

Submission Instructions

- You must form groups of 2 to 3 people
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

  `Exercise_09_MatriculationNumber1_MatriculationNumber2.zip`
- Provide in the archive:
  i. your code, accompanied with sufficient comments
  ii. a PDF report with answers, solutions, plots and brief instructions on executing your code
  iii. a README file with the group member names, matriculation numbers and emails
  iv. Data necessary to reproduce your results
- The subject of your submission mail **must** contain the string [SNLP] (including the braces) and explicitly denoting that it is an exercise submission, e.g:

  `[SNLP] Exercise# Submission MatriculationNumber1 MatriculationNumber2`
- Depending on your tutorial group, please send your assignment to the **corresponding tutor**:
  * Mo. 16-18: Lukas Lange *s9lslang@stud.uni-saarland.de*
  * Th. 14-16: Harshita Jhavar *snlp18.thursday@gmail.com*
  * Fr. 10-12: Marimuthu Kalimuthu *mkalimuthu@lsv.uni-saarland.de*
- Note that if you submit to any other emails than the above, it will not be graded at any cost.
- **Copying solutions is strictly forbidden**. If any form of plagiarism is found out, then all teams involved in the practice would get **0 points**. If this happens again, then they will be excluded from the course. Note that copying from web resources also counts as plagiarism, unless it is properly cited.