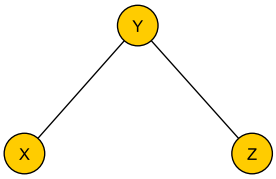
Statistical Natural Language Processing
Assignment 9

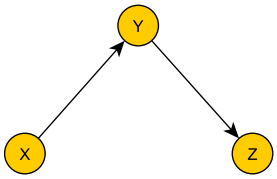
Sven Stauden 2549696
Janis Landwehr 2547715
Carsten Klaus 2554140

Full points!

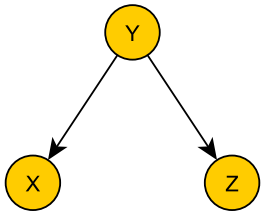
Exercise 1



(a) MRF of the stated problem



(b)



(c)

The given problem description **does not** imply $p(x|z) = p(x|z,y)$. We could adapt the energy functions g and f so that $p(x|z) = p(x|z,y)$ holds by setting them to

$$h(a,b) = P(a) \cdot P(b|a)$$

$$g(a,b) = P(b|a)$$

Doing so, we have implemented a Bayesian Network described in (b). Choosing the following energy functions, which implement the Bayesian Network (c)

$$h(a,b) = P(b) \cdot P(a|b)$$

$$g(a,b) = P(b|a)$$

, we have a counterexample for which the assumption $p(x|z) = p(x|z,y)$ does not hold because here X provides a dependency to Y so $p(x|y,z) \neq p(x) = p(x|z)$.

Exercise 2

2.1

- numerous value between 1000 and 2050 - useful/necessary to identify annual dates. Numbers not in this range are likely to have a different meaning.
- information about the document type - Conversational texts like emails have a different sentence structure than non-addressed text like in a science paper. Knowing the context allows to apply different entity strategies.
- n-gram frequency - tokens that often appear together might also set up a common entity. E.g. "HP LaserJet 4200". Instead of setting the entities to each word individually ("HP" - Company, "LaserJet"- ?, "4200"-Number) one can specify special occurences ("HP LaserJet 4200"- Product).
- Capitalization - Words with capital letters which are not located in the beginning of a sentence are very likely to be names of persons, locations, companies, etc. These words are worth considering in a specialized dictionary managing names.
- trigger words - some words allow to indicate the entity of neighboring words with a more or less good predictability. E.g. "drinking tea", "drinking coffee", "drinking ...".

2.2

- CRFs are undirected while HMM are always directed. This means that HMM only consider dependencies in one direction. CRFs define their dependencies over the weights. For HMMs only the dependency **from** the previous token is relevant.
- While HMM only consider the current and the previous label, CRFs take all labels of the sequence into account and therefore are able to consider more global features.

Exercise 3

3.1. - 3.3

See python implementation

The entire implementation which got also submitted is located in this repository: `https://github.com/Janis90/snlp/tree/master/HW09`

Out of this repository, the model file can be downloaded from here: `https://github.com/Janis90/snlp/blob/master/HW09/model`

3.4

```
processed 51578 tokens with 5942 phrases; found: 5870 phrases; correct: 4756.
accuracy:  96.74%; precision:  81.02%; recall:  80.04%; FB1:  80.53
              LOC: precision:  84.21%; recall:  84.76%; FB1:  84.48  1849
             MISC: precision:  86.65%; recall:  67.57%; FB1:  75.93  719
              ORG: precision:  73.93%; recall:  73.15%; FB1:  73.54  1327
PER: precision: 80.76%; recall: 86.59%; FB1: 83.57 1975
```