Machine Language Processing, Ling 331
Wintersemester 2014
University of Konstanz
Miriam Butt

Homework 1

(1) and (2) provide an English and a German text, respectively. You can decide to work with either text.

(1) At the age of one, Harry had somehow survived a curse from the greatest dark sorcerer of all time, Lord Voldemort, whose name most witches and wizards still feared to speak. Harry's parents had died in Voldemort's attack, but Harry had escaped with his lightning scar, and somehow — nobody understood why — Voldemort's powers had been destroyed the instant he had failed to kill Harry.
[J.K. Rowling, *Harry Potter and the Chamber of Secrets*]

(2) Damals im Alter von einem Jahr, überlebte Harry auf merkwürdige Weise den Todesfluch des größten schwarzen Magiers aller Zeiten. Die meisten Hexen und Zauberer hatten immer noch Angst, dessen Namen auszusprechen: Lord Voldemort. Harrys Eltern starben bei Voldemorts Überfall, doch Harry kam mit der blitzförmigen Narbe davon. Voldemorts Macht jedoch fiel in eben jenem Augenblick in sich zusammen, als es ihm misslungen war, Harry zu töten. Und keiner konnte das begreifen. [J.K. Rowling, *Harry Potter und die Kammer des Schreckens*]

Imagine that your ultimate goal is to write a Question-Answer system on the lines of IBM's Watson. Assume that the texts in (1) or (2) are your knowledge base and that your system should be able to answer the following:

- What did Harry survive?/Was hat Harry überlebt?
- Does Lord Voldemort still have his powers?/Hat Lord Voldemort noch seine Macht?
- Who is the greatest sorcerer of all time?/Wer ist der größte schwarze Magier aller Zeiten?
- Who was Harry attacked by?/Von wem wurde Harry überfallen?

In the exercises that are part of the course, we will go through each step needed to automatically process the above texts so that a Question-Answer system could in principle be built.

1 Tokenization

A first step is *Tokenization*.

1. What is the goal of tokenization?
2. How does tokenization work?
3. Go to the Xerox Linguistic Tools Website and let their system tokenize your text. Check whether the results are what you would expect and discuss the overall results.
4. Hand in the resulting tokenization along with your answers.

2 POS Tagging

1. What is the goal of POS Tagging?
2. How does POS Tagging work?
3. Have your Text be POS-tagged by one of the available on-line POS-taggers.
4. Discuss the results (are they what you would have expected, if not, why not).
5. Hand in the POS tagged text along with your answers.

3 Morphological Analysis

1. Why is a morphological analysis important/useful in automatic language processing? (You just need to provide a short answer.)
2. Explain some advantages and disadvantages of a "Porter Stemmer".

3. Use the on-line Porter Stemmer Demo on this website to analyze your text:

   `http://text-processing.com/demo/stem/`

   What works well, what doesn't?

4. Now run your text through the Xerox tools. What kind of morphological analysis is being performed here and what is the underlying technology?

5. Discuss the results (are they what you would have expected, if not, why not).

4 Parsing

1. Assume you had a grammar with the context-free rules and lexical entries specified in in (6) or in (7).

   (a) Which of the following sentences can the grammar parse? Which will it fail on? Note that you can ignore punctuation and capitalization for now (that will need to be taken care of by a tokenizer).

   German:

   (3) a. Der kleine Hund sieht die Katze.
   b. Ein Hund bellt.
   c. Hunde bellen.
   d. Die kleine, schnelle Katze miaut.
   e. Der Hund in dem Garten sieht die Katze.
   f. Der Hund sieht die Katze auf der Mauer neben dem Haus.

   English:

   (4) a. The small dog sees the cat.
   b. A dog barks.
   c. Dogs bark.
   d. The small, quick cat meows.
   e. The dog in the garden saw the cat.
   f. The dog saw the cat on the wall beside the house.

   (b) What changes need to be made to the grammar so that all the sentences can be parsed?

2. What does the Kleene `*` express? And the Kleene `+`?

3. Show how the sentences in (5) can be parsed via the parsing strategies in (a) and (b).

   (5) The bird checks for worms.

   (a) Bottom-up, Left-to-Right

   (b) Top-down, Right-to-Left

(6) **Deutsch**

S ⟶ NP VP
NP ⟶ D Adj* N
VP ⟶ V NP (PP)
PP ⟶ P NP

der	D
die	D
dem	D
den	D
ein	D
eine	D
Hund	N
Hunde	N
Katze	N
Mauer	N
Garten	N
Segel	N
Haus	N
Schiff	N
schnelle	Adj
kleine	Adj
jagt	V
sieht	V
bellt	V
miaut	V
flattern	V
segel	V
in	P
auf	P
neben	P

(7) **English**

S ⟶ NP VP

NP ⟶ D Adj* N

VP ⟶ V NP (PP)

PP ⟶ P NP

the	D
a	D
dog	N
dogs	N
cat	N
wall	N
garden	N
house	N
sails	N
wind	N
quick	Adj
small	Adj
chases	V
flutter	V
sails	V
saw	V
barks	V
meows	V
sails	V
in	P
on	P
beside	P