SNLP – Statistical Natural Language Processing

1. SNLP - Motivationen

SNLP-Systeme sind
- **Schnell** und **billig** zu produzieren
- Erlauben schnelles „prototyping“
- **Robust**
  - Egal wie schlecht der Input ist, es gibt immer einen Output

1. SNLP - Motivationen

- SNLP-Systeme liefern gute Ergebnisse („real-world“ applications)
- Ambiguität von Sprache
- Wahrscheinlichkeitsstrategien wesentlicher Bestandteil menschlicher Kognition
- SNLP-Systeme werden dem „kreativen Charakter“ von Sprache gerecht

2. SNLP - Voraussetzungen

SNLP-Systeme bedienen sich der **Techniken maschinellen Lernens**
- Benötigen (annotierte) Textkorpora zu Trainings- und Testzwecken
- Z. B. Penn Treebank, Switchboard,...

2. SNLP - Korpora

Einige statistische „Spielereien“ im Textkorpus
- Zählen der häufigsten Wörter

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in *Tom Sawyer*.

2. SNLP - Korpora

Einige statistische „Spielereien“ im Textkorpus

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11–50	540
51–100	99
> 100	102

Table 1.2 Frequency of frequencies of word types in *Tom Sawyer*.

2. SNLP - Korpora

Erkenntnisse:

- ungleiche Verteilung der Wort-“types“
- Einige wenige Wörter kommen sehr häufig vor
- 90% tauchen weniger als zehn mal auf, davon 12% weniger als drei mal
- => Problem für SNLP: verwendete Korpora müssen sehr groß sein

3. SNLP – Was ist das?

- Statistische Verfahren zunächst in der Spracherkennung eingesetzt => Ausdehnung auf andere NLP-Bereiche
- Grundlagen der SNLP: Wahrscheinlichkeitstheorie & Informationstheorie („noisy channel model“)

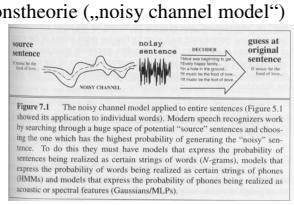

Figure 7.1 The noisy channel model applied to entire sentences (Figure 5.1 showed its application to individual words). Modern speech recognizers work by searching through a huge space of potential “source” sentences and choosing the one which has the highest probability of generating the “noisy” sentence. To do this they must have models that express the probability of sentences being realized as certain strings of words (*N*-grams), models that express the probability of words being realized as certain strings of phones (HMMs) and models that express the probability of phones being realized as acoustic or spectral features (Gaussians/MLPs).

3. SNLP – Beispiel

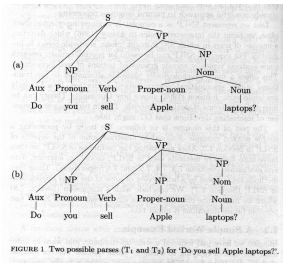

FIGURE 1 Two possible parses (T_1 and T_2) for ‘Do you sell Apple laptops?’.

3. SNLP – Beispiel

$A \rightarrow \beta \; [p]$

A simple grammar is shown in Figure 2 (adapted from Jurafsky and Martin (2000)).

S → NP VP	.80	Det → that	.05
S → Aux NP VP	.15	Det → the	.80
S → VP	.05	Det → a	.15
NP → Det Noun	.20	Noun → laptops	.50
NP → Proper-Noun	.35	Noun → desktops	.50
NP → Nom	.05	Verb → sell	.30
NP → Pronoun	.40	Verb → ship	.55
Nom → Noun	.75	Verb → repair	.15
Nom → Noun Noun	.20	Aux → can	.60
Nom → Proper-Noun Noun	.05	Aux → do	.40
VP → Verb	.55	Proper-Noun → Apple	.50
VP → Verb NP	.40	Proper-Noun → Sony	.50
VP → Verb NP NP	.05	Pronoun → you	.40
		Pronoun → I	.60

FIGURE 2 A very simple probabilistic grammar for English

3. SNLP – Beispiel

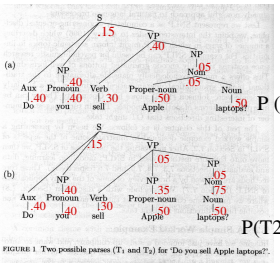

FIGURE 1 Two possible parses (T_1 and T_2) for ‘Do you sell Apple laptops?’.

P (T1) = 7.2 *10 exp-7

P(T2) = 4.725 *10 exp-7

4.1 SNLP – Wie geht das?

Modeling

- „Modeling is the general task of constructing machinery which mimics some task“

1. Struktur des Modells (z. B. Syntax-Modell => grammatikalische Regeln)
2. Parametrisierung des Modells (Variablen, die bestimmte Wahrscheinlichkeiten tragen können)
3. Instantiierung des Modells (d.h. Parameter bekommen Wahrscheinlichkeitswerte zugewiesen) => über Lernen/Schätzen

4.1 SNLP – Modeling

Problem bei Modeling:

- Overfitting (Modell ist zu speziell)
- Underfitting (Modell ist zu allgemein)

Zwei verschiedene Modelle:

- Generative vs. discriminative models

4.2 SNLP – Estimation/Learning

Parameter eines Models müssen aufgrund von Trainingsdaten Werte zugewiesen bekommen.

Dabei ist zu beachten:

- **Bias** (Verzerrungen der Daten durch):
  - Bias by design (Occam ´s razor)
  - Bias in design (maximum likelihood estimators, maximum entropy models)
- **Scalability**
  - Behandelt „trade-off" zwischen Anzahl der Parameter und Computerressourcen

4.2 Estimation: Maximum likelihood (MLE)

- Bestimmung der Parameter des Modells so, dass die Wahrscheinlichkeit auf den Trainingsdaten am höchsten ist.
- einfache (inkorrekte) Bestimmung durch Zählen
  - z. B. bei „part-of-speech tagging": (Wie oft bekommt ein Wort einen bestimmten „tag" zugewiesen) / (Anzahl der verschiedenen möglichen „tags" für dieses Wort)

4.2 Maximum likelihood - smoothing

- MLE is biased towards the training set => Tendenz zum „overfitting"
- Smoothing ist eine Technik, die Daten so zu verändern, dass die Perfomance der SNLP-Anwendung steigt

4.2 Estimation – Maximum Entropy (Maxent)

- Maß des Informationsgehaltes
- Kann z.B. genutzt werden, als Maß, wie viel Information in einer best. Grammatik steckt oder wie gut eine best. Grammatik eine Sprache abbildet,...
- „Maxent" Modelle können Dependenzen zwischen Parametern (Wörtern) modellieren.

MLE und Maxent Modelle sind in eine Art „Zusammenfassung" der Trainingsdaten.

4.2 Estimation – Weitere Ansätze

- Kernel Methoden
- Ensemble Approaches

4.3 SNLP - Evaluation

Which method is best for a given task?

⇒ **Quantitative comparison between different approaches**.

Zwei Maße (e.g. text classification):

- **Precision:** number of correctly classified documents/number of classifications posited
- **Recall:** number of documents assigned a particular classification / number of documents in the testing setwhich actually belong to that class

4.3 SNLP - Evaluation

- training & testing set
- Upper bound upon performance: human performance
- Lower bound upon performance: chance

5. SNLP - Applications

- Part-of-Speech Tagging
- Statistical Parsing (shallow parsing)
- Text Classification
- Question answering (intelligent information retrieval)
- Machine Translation
- OCR/Speech Recognition
- Augmentive communication systems