Schlussprüfung Einführung in die Computerlinguistik I HS 10

Aufgabenstellung: Simon Clematide

Prüfung vom 20. Januar 2011
Institut für Computerlinguistik, Universität Zürich

Vorname _____ Matrikelnummer _____

Nachname _____

Für Studierende der folgenden Studiengänge:

☐ BA - Studiengang Computerlinguistik (Phil. Fakultät)

☐ BA - Studiengang Computerlinguistik und Sprachtechnologie (Phil. Fakultät)

☐ BA-Studierende (Wirtschaftswissenschaftliche Fakultät)

☐ Studierende des Nebenfachs Informatik mit Studienbeginn ab WS 04/05

☐ Multidisziplinfach (ETH)

☐ Andere:

Nur für Lizentiatsstudierende der Computerlinguistik als ein Fach aus der Phil. Fakultät:

Strasse: _____ Hauptfach: _____

PLZ/Ort: _____ E-Mail: _____

Die Prüfungsresultate von ECL und PCL von den Lizentiatsstudierenden werden (zusammen) per Post verschickt. *Bitte Adresse nicht vergessen!*

Aufgabe Nr.:	1	2	3	4	5	6	7	8	**Summe**
Punktzahl:	24	8	8	12	12	10	10	6	90
Davon erreicht:									

Note SU: _____ Note SP: _____

Endnote: _____ Bestanden: ☐ Ja ☐ Nein

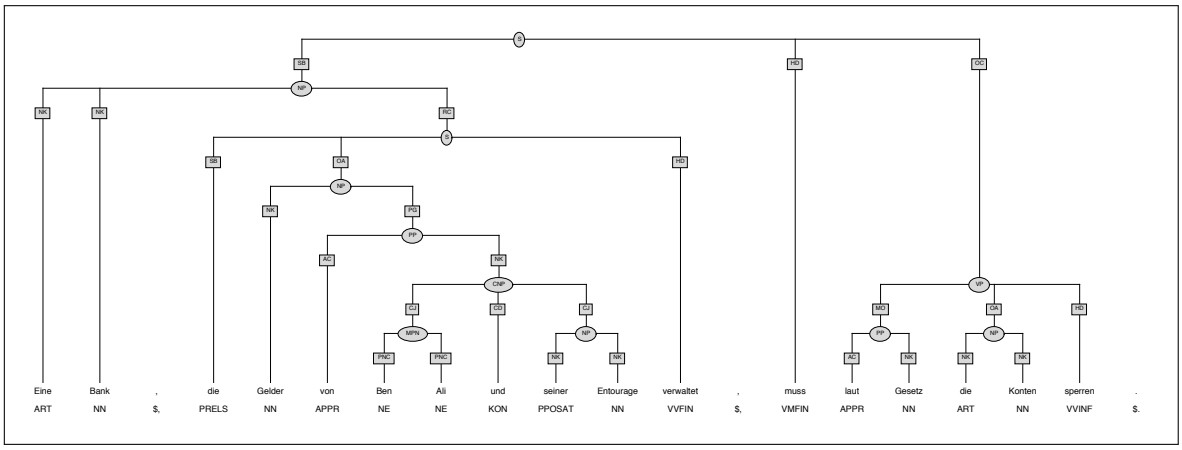
Auf jedes zusätzliche Blatt mit Lösungen den Nachnamen schreiben! Viel Erfolg!

Wichtige Hinweise

Punkte-Maximum: 90 (pro Minute 1 Punkt)
Hinweis: Bitte schreiben Sie in einem überlegten und knappen, aber verbalen Stil (keine Stichwortsammlungen). Bei inhaltlichen Auswahlsendungen, wo einfach mal alles spontan hingeschrieben wird und Falsches wie Korrektes munter vermischt sind, behalte ich mir Abzüge vor. Erlaubtes Hilfsmittel ist die Referenzkarte zum Annotieren.

1. Morphologische und syntaktische Annotation (24 Punkte)

18 (a) Zeichnen Sie den TIGER-Baum im Stil des Annotate-Werkzeugs mit **Wortart** (STTS **ohne** Morphologieanalyse), **Phrasenkategorien** und **syntaktischen Funktionen** zu folgendem Satz (auf der letzten Prüfungsseite ist der Satz bereits im Querformat aufgedruckt):
Eine Bank, die Gelder von Ben Ali und seiner Entourage verwaltet, muss laut Gesetz die Konten sperren.



3 (b) Wo liegt für Sie die syntaktische Hauptschwierigkeit im obigen Satz? Begründen Sie kurz.

Es gibt einige Dinge, welche nicht ganz trivial sind.

- Koordination innerhalb einer PP
- Relativsatz
- Anschluss des Modifikators „laut Gesetz“
- Anschluss von „von Ben Ali und seiner Entourage“

1 (c) Geben Sie für die folgenden Wortformen aus obigem Satz die detaillierte **morphologische Analyse** nach STTS-Schema:

a) seiner: Dat.Sg.Fem

b) Gesetz: Dat.Sg.Neut

2 (d) Welche Wortarten sind für die Wortform „laut“ im Deutschen denkbar? Begründen Sie linguistisch (!), welche Wortart im obigen Satz vorliegt.

- Präposition: Semantische Begründung: Es wird im Satz ja im Sinn von „gemäss“ verwendet, nicht im Sinn von „lärmend“; Syntaktische Begründung: „Gesetz“ ist abhängig von „laut“

- Adjektiv (adverbial, prädikativ, aber nicht attributiv, weil ja keine Flexionsendung da ist)
- Adverb (Wenn man den Annotationskonventionen von STTS folgt, muss das adverbial verwendete „laut“ als Adjektiv aufgefasst werden. Unabhängig von STTS ist „laut“ als Adverb durchaus möglich.)

2. Spam-Filter (8 Punkte) In einem Testset von 1000 E-Mails hat es 400 echte Spam-Mails.

4 (a) Student Bruno hat einen Spam-Filter programmiert, der Spam mit einer Precision von 80% erkennt. Wieviele False-Positives (FP) entstehen, wenn sein Filter insgesamt 300 Positive ergibt? 60%

4 (b) Studentin Gerda hat ihren selbst programmierten Filter so eingestellt, dass er gleich viele Positive wie Negative erzeugt. Wie hoch ist ihre Precision, wenn sie einen Recall von 90% hat? 72%

8 **3. Multilingualität (8 Punkte)** Inwiefern ist Multilingualität bereits für einfache Text-Segmentierung ein Problem? Wie lässt es sich lösen?

- Tokenisierung benötigt bereits Sprachidentifikation, da sie sprachspezifisch ist (Abkürzungswörterbücher, Interpunktionskonventionen)
- Ansätze zur Sprachidentifikation: Siehe Vorlesungsunterlagen dazu!

4. Evaluation (12 Punkte)

6 (a) Worum handelt es sich beim BLEU-Score? Was wird damit gemessen und anhand wovon?

Siehe Unterlagen.

6 (b) Wieso sind Evaluationen in der CL im Allgemeinen wichtig? Wozu dienen sie?

- Systemoptimierung: Messen von Fortschritt bzw. Rückschritt bei Änderungen
- System- und Methodenvergleich: Welche Ansätze erledigen eine Aufgabe besser?
- Operationalisierbare (objektive) Kriterien und Tests, welche Qualitätsaussagen nachprüfbar machen.
- Evaluationen an Goldstandards erlauben systematische Fehleranalysen

12 **5. Maschinelle Übersetzung (12 Punkte)**

Vergleichen Sie den transferbasierten Ansatz mit dem statistischen. Welche wichtigen Vorteile bzw. Nachteile gibt es?

- Transferbasierter Ansatz: Syntaxanalyse, Syntaxgenerierung und Transferkomponente von Hand implementieren. Aufwändig herzustellen. Normalerweise lässt sich dadurch die syntaktische Korrektheit der Übersetzung besser gewährleisten. Benötigt keine Trainingsdaten. Die Differenz zwischen Quell- und Zielsprache kann sehr gross sein.
- Statistischer Ansatz: Keine echte Syntaxanalyse und -generierung. Usuelle Sprachverwendung werden gut umgesetzt. Leicht herzustellen mit Standard-Software, falls genügend (parallele) Trainingsdaten vorhanden sind (Sprachmodelle und Übersetzungsmodelle). Für die Übersetzung zwischen Sprachen mit geringen Unterschieden besser geeignet als für Sprachen mit sehr unterschiedlicher Typologie und Sprachbau.

10 **6. Sprachsynthese (10 Punkte)** Welche Verarbeitungsstufen sind typisch für ein modernes Sprachsynthesesystem wie Mary?

Siehe Unterlagen!

7. Kontextfreie Regel-Grammatiken (10 Punkte)

Gegeben sei eine englische Mini-Grammatik $G = \langle\{\mathrm{IP}, \mathrm{I}', \mathrm{VP}, \mathrm{NP}, \mathrm{V_i}, \mathrm{PN}, \mathrm{N}, \mathrm{D}\}, L, R_{syn} \cup R_{lex}, \mathrm{IP}\rangle$ im Stil der IP-Analysen (IP=inflectional phrase, d.h. Phrase mit flektiertem (Hilfs-)Verb als Kopf):

Syntaxregeln: $R_{syn} = \left\{ \begin{array}{llll} \mathrm{IP} \to \mathrm{NP\ I}', & \mathrm{I}' \to \mathrm{I\ VP}, & \mathrm{VP} \to \mathrm{V_i}, & \mathrm{VP} \to \mathrm{V_t\ NP}, \\ \mathrm{NP} \to \mathrm{D\ N}, & \mathrm{NP} \to \mathrm{PN}, & & \end{array} \right\}$

Lexikonregeln: $R_{lex} = \left\{ \begin{array}{lllll} \mathrm{I} \to \mathrm{might}, & \mathrm{I} \to \mathrm{should}, & \mathrm{V_i} \to \mathrm{sleep}, & \mathrm{V_t} \to \mathrm{meet}, & \mathrm{V_i} \to \mathrm{hit}, \\ \mathrm{D} \to \mathrm{the}, & \mathrm{D} \to \mathrm{her}, & \mathrm{PN} \to \mathrm{Lucy}, & \mathrm{N} \to \mathrm{friend}, & \mathrm{N} \to \mathrm{dog}, \end{array} \right\}$

3 (a) Schreiben Sie die Linksableitung für den Satz „Lucy should meet her friend“ in der Form: IP $\Rightarrow \ldots \Rightarrow$Lucy should meet her friend

IP $\Rightarrow$NP I′ $\Rightarrow$PN I′ $\Rightarrow$Lucy I′ $\Rightarrow$Lucy I VP $\Rightarrow$Lucy should VP $\Rightarrow$Lucy should V_t NP $\Rightarrow$Lucy should meet NP $\Rightarrow$Lucy should meet D N $\Rightarrow$Lucy should meet her N $\Rightarrow$Lucy should meet her friend

7 (b) Ergänzen Sie die Grammatikregeln und das Lexikon, damit auch Sätze erkannt werden mit satzwertigen Objekten wie: „Lucy might say that her friend should meet“ oder „her friend might think that Lucy should say that her dog should sleep“, aber nicht *„that her friend might think“.

Lexikonregeln:	$\mathrm{C} \to \mathrm{that}$	$\mathrm{V_s} \to \mathrm{say}$	$\mathrm{V_s} \to \mathrm{think}$
Syntaxregeln:	$\mathrm{VP} \to \mathrm{V_s\ CP}$		$\mathrm{CP} \to \mathrm{C\ IP}$

8. Merkmalstrukturen (6 Punkte)

$M_1 = \begin{bmatrix} \mathrm{NUM} & \mathrm{sg} \end{bmatrix}$ $M_2 = \begin{bmatrix} \mathrm{AGR} & \begin{bmatrix} \mathrm{NUM} & \mathrm{pl} \end{bmatrix} \end{bmatrix}$ $M_3 = \begin{bmatrix} \mathrm{AGR} & [\,] \end{bmatrix}$

1 (a) Können die Merkmalstrukturen M_1 die Merkmalstruktur M_2 unifiziert werden?
☐ **Ja** ☐ Nein ☐ Undefiniert

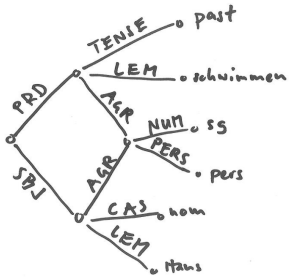
1 (b) Können die Merkmalstrukturen M_2 die Merkmalstruktur M_3 unifiziert werden?
☐ **Ja** ☐ Nein ☐ Undefiniert

4

(c) Zeichnen Sie die folgende koreferente XLE-Merkmalstruktur als Graphen auf:

```
"Hans schwamm"

      [PRD  3[AGR [NUM sg, PER 3]
             LEM schwimmen, TENSE past]
  7[SBJ  1[AGR [3-AGR]
           CAS nom, LEM Hans]                ]
```



(Bitte Baum mit STTS-Wortartentags, Phrasen und Funktionen, aber ohne Morphologie zeichnen!)