FEATUREAUSWAHL UND OPTIMIERUNG AM BEISPIEL SPRACHENIDENTIFIZIERUNG

SEMINAR: KLASSIFIKATION

DOZENT: STEFAN LANGER

CIS, UNIVERSITÄT MÜNCHEN

Wozu Sprachenerkennung

Interne Verarbeitung

- Sprachenerkennung zur weiteren linguistischen Verarbeitung
  - Tokenisierung, Lemmatisierung (Stemming) etc etc
- + Kodierungserkennung um überhaupt mit einem Text arbeiten zu können

Filter

- Wähle Dokumente nur aus einer bestimmten Sprache aus
- Finde Teildokumente bei mehrsprachigen Dokumenten

Sprachenidentifizierung

ویلیام شکسپیر در ۲۶ آوریل سال ۱۵۶۴ در انگلستان در شهر استراتفورد متولد شد. شهرت شکسپیر به عنوان شاعر، نویسنده، بازیگر و نمایشنامه نویس منحصربه‌فرد است و برخی او را بزرگ ترین نمایشنامه نویس تاریخ می‌دانند، اما بسیاری از حقایق زندگی او مبهم است.

fa

Medan Brand var eit drama Ibsen sleit lenge med, kom Peer Gynt omntrent av seg sjølv. Dramaene står i et refleksjonstilhøve til kvarandre

nn

Drama radio enwog gan Dylan Thomas a gyhoeddwyd yn 1954 yw Under Milk Wood. Mae'r ddrama yn disgrifio digwyddiadau mewn un diwrnod yn unig, yn y pentref dychmygol Llareggub, er cred llawer fod nifer o'r cymeriadau yn seiliedig ar bobl go iawn ag oedd yn byw yn Nhalacharn.

cy

Sprachenerkennung – Beispiel

(nach Dunning: Statistical Identification of Language, 1994):

e pruebas biquimica

man immunodeficiency

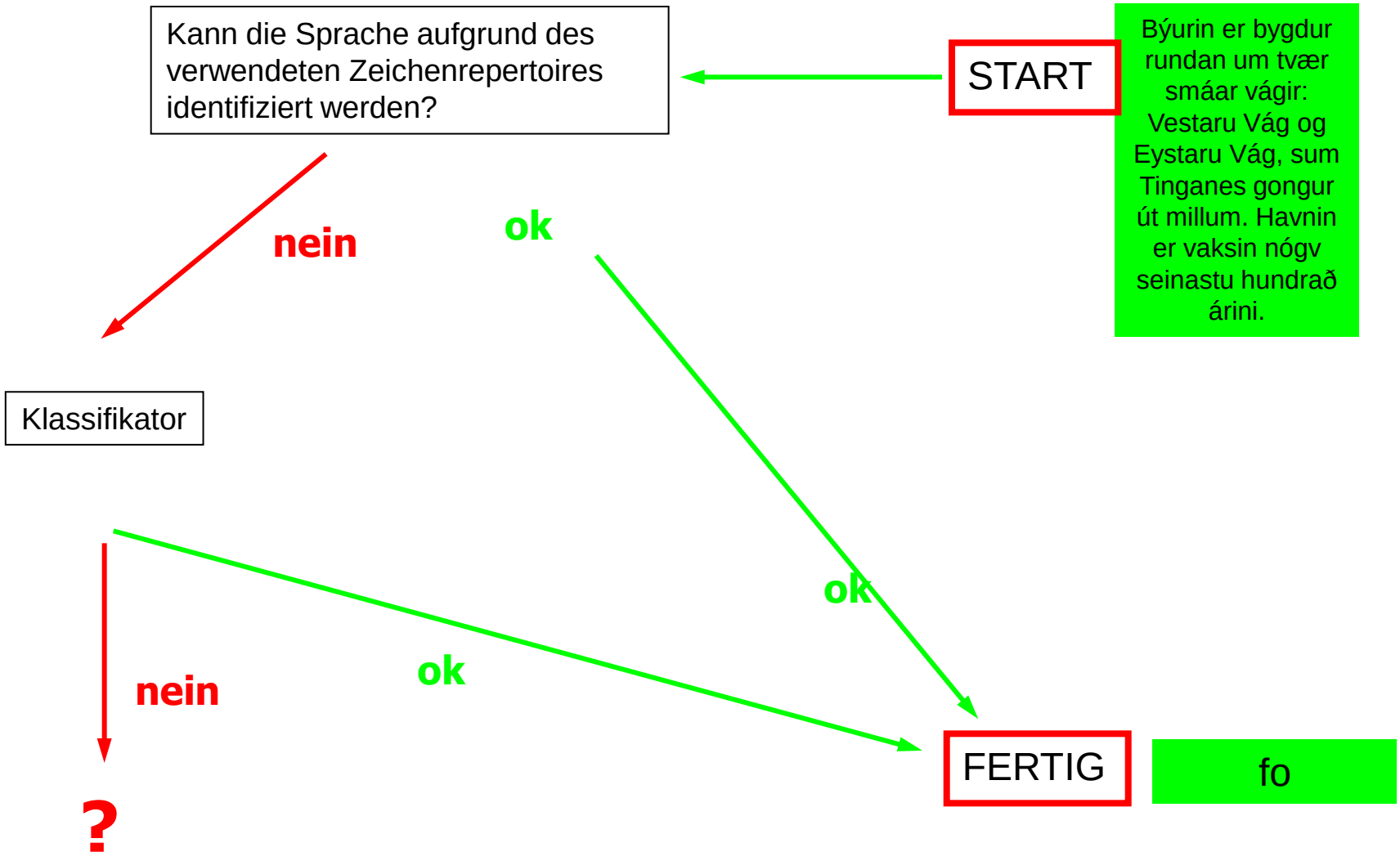
faits se sont produi

er biochemischen Forsch

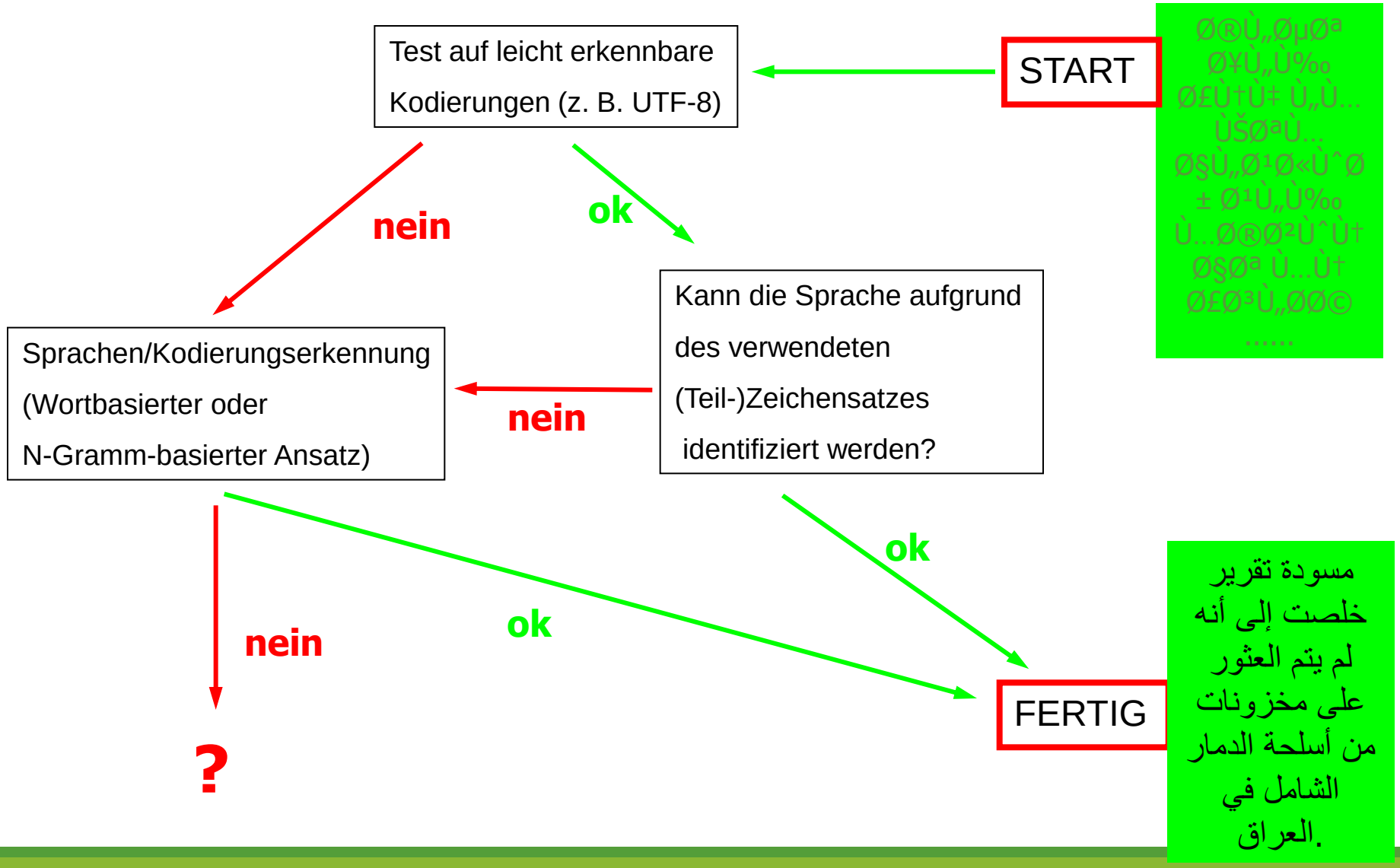
Sprachenerkennung – Features

- Semantische Analyse nicht notwendig
  - Mapping auf Embeddings nicht zielführend
- Wörter oder kleinere Einheiten ausreichend

Algorithmus Sprachenerkennung – zweistufig



Algorithmus Sprachen- und Kodierungserkennung



Featureauswahl: Wortbasierte Erkennung

Features:

- Subset des Wortschatzes der Trainingsdokumente (abhängig vom zu klassifizierenden Dokumenttyp und dem morphologischen System einer Sprache)

- Naiver Algorithmus: vergleiche Wörter im Dokument mit Wörtern im Wörterbuch / Modell
- Erkennungswert eines Wortes abhängig von:
  - Worthäufigkeit (z.B. naïve Bayes)
  - Eindeutigkeit
  - Länge

N-Gramm-Features

digwyddiadau mewn → [dig, igw, gwy, wyd, ydd, ddi, dia]

Daten

- Für jede Sprache:
  - Buchstaben N-Gramm-Liste mit TF-IDF/Häufigkeit
  - N = 2-4

Vergleich der Featureauswahl

Wortbasiert	**N-Gramm-Ansatz**
Trainingskorpus muss nicht ganz sauber sein, da manuelle Überprüfung möglich	Sauberes Trainingskorpus
Aufwändiges Training, wenn manuell überprüft	Training einfach
Nachträgliche Überprüfung und Korrektur unproblematisch	Nachträgliche Überprüfung / Revision kaum möglich, außer über Trainingskorpus
relative große Datenbasis zur Erkennung	kleine Datenbasis
Neue Kodierungen einfach zu ergänzen	Konversion des Trainingskorpus nötig zur Ergänzung von neuen Kodierungen
Schlecht geeignet für sehr kurze Dokumente oder Listen seltener Wörter	Auch für sehr kurze Dokumente geeignet
Nicht für Sprachen ohne durch Leerzeichen markierte Wortgrenzen (Japanisch, Chinesisch...	Alle Sprachen

Sprachenidentifizierung - Daten

- Datenquellen
  - Wikipedia
  - Wictionary
  - Alle Datenquellen mit monolingualen Dokumenten

Datenaufbereitung:

Parsen der Dokumente und Erstellung von Wort-/N-Grammlisten

Größe der Features: Textabdeckung als Kriterium

Fehlerquellen

Fremdsprachige Zitate

Wörter aus anderen Sprachen (insbesondere bei Sprachen aus Ländern mit weiterer, größerer Verkehrssprache, z.B. Catalan -> Spanisch)

Text mit zahlreichen Eigennamen

Kein Fließtext sondern Listen (z.B. Produktlisten, Namenslisten)

Algorithmen: mögliches Finetuning

Wie oft wird ein Wort/N-Gramm berücksichtigt?

Wortlänge

Aussortieren unpassender Einträge bei Wortliste

- Andere Sprachen

- Anderer Zeichensatz

Bei manuelle Featureoptimierung:

- Text -> Wortliste (optimiert) -> N-Gramme

Schwierige Fälle

Corpora:

Sehr kurze Texte

Mehrsprachige Texte

Text mit zahlreichen Eigennamen

Kein Fließtext sondern Listen (z.B. Produktlisten, Namenslisten)

Sprachen:

Nah verwandte Sprachen

Untersprachen (z.B. BE, AE)

Primär politisch motivierte Sprachentrennung