Sprachenerkennungssystem
Pflichtenheft

Version 1.0

Mark Schmid

27. Juni 2003

Inhaltsverzeichnis

1 Einleitung

1.1 Umfeld

Für fortgeschrittene Suchmaschinen oder Agenten ist es nötig rasch zu entscheiden, in welcher Sprache ein gegebener Text verfasst ist. Weiterentwicklungen könnten sogar soweit gehen, dass zusätzlich entschieden werden kann, welcher Wissensgemeinschaft ein bestimmter Text zuzuordnen ist. Grundsätzlich unterscheiden sich Sprachen und in geringerem Mass Wissensgemeinschaften in der Häufigkeitsverteilung von n-Grammen.

Natürliche Sprachen können sogar schon durch die Häufigkeitsverteilung der Buchstaben unterschieden werden. Dazu ist aber eine hinreichend grosse Statistik nötig. Erwünscht ist eine Sprachenerkennung mit möglichst wenig Text. Je mehr Information über die syntaktische Struktur von n-Grammen in einer Sprache ausgenutzt wird, desto weniger Text wird für die Erkennung notwendig sein.

1.2 Aufgabenstellung

Die offizielle Aufgabenstellung kann unter [1] eingesehen werden.

1.3 Thema

Diese Diplomarbeit setzt sich hauptsächlich mit folgenden Themenbereichen auseinander

- Neuronale Netze
- Wahrscheinlichkeitsrechnung und Statistik
- Planung, Design und Implementation von Software
- Experimentelle Tätigkeit

1.4 Ziele

1.4.1 Auszug aus der Aufgabenstellung

”Mit Hilfe eines neuronalen Netzes ist ein Sprachenerkennungssystem zu bauen, das die wichtigsten indogermanischen (europäischen Sprachen) zuverlässig mit möglichst kurzen Textsequenzen erkennen kann.“

1.4.2 Erforschung der Eignung Neuronaler Netze für Sprachenerkennung

Kategorie *Pflichtziel*

Detailbeschrieb Im Rahmen dieser Erforschung soll analysiert werden, inwiefern sich die Verwendung eines Neuronalen Netzes zur Ermittlung der natürlichen Sprache, in der ein Text verfasst ist, eignet, insbesondere hinsichtlich der Frage, ob mit wenig Eingabetext gute Resultate erzielt werden können.

Als Massstab bezüglich der Kürze des Eingabetextes soll, wie in der Aufgabenstellung erwähnt, das Sprachenerkennungswerkzeug "Language Identifier" von LexTek [2] herangezogen werden. Dieses braucht mindestens 200 Zeichen als Eingabetext.

Werkzeuge Die zum Erreichen dieses Ziels primär eingesetzten Werkzeuge sind:

- Matlab [3] / Mathematica [4]
- Programmiersprachen wie z.B. Java [5]

Abgrenzung Ein allfälliger Befund, dass sich Neuronale Netze nicht oder nur bedingt für die erwähnte Aufgabe eignen, ist explizit zulässig, sofern diese Aussage im Rahmen der Forschungsarbeit aufgezeigt und gestützt werden kann.

1.4.3 Erstellung einer Sprachenerkennungs-Applikation

Kategorie *Pflichtziel*

Detailbeschrieb Es soll eine auf einem geeigneten Neuronalen Netz basierende Applikation entwickelt werden, die es dem Benutzer erlaubt, einen Text einzugeben und für den eingegebenen Text einen Wahrscheinlichkeitsvektor für die Sprache, in welcher der Text verfasst ist, ausgibt. Dabei sollen mindestens folgende Sprachen unterstützt sein:

- Deutsch
- Englisch
- Französisch
- Italienisch
- Spanisch

Optional kann die Unterstützung von weiteren Sprachen implementiert werden.

Werkzeuge Die Applikation wird in der Programmiersprache Java [5] erstellt. Einzelne Komponenten können aber auch in anderen Programmiersprachen entwickelt werden, wenn diese dafür geeigneter erscheinen.

Abgrenzung Die Applikation soll in Kommandozeilen-orientierter Form vorliegen. Optional kann diese auch um ein GUI- oder Web-Frontend erweitert werden.

Folgende Einschränkungen bestehen:

- Orthographisch korrekter Eingabetext
- Der Eingabetext besteht aus in der entsprechenden Sprache verwendeten Zeichen (also z.B. im Deutschen “ö” statt “oe”)

1.4.4 Weiterführende Erforschung

Kategorie *optionales Ziel*

Detailbeschrieb Gemäss Aufgabenstellung könnte versucht werden, einen Text maschinell einer bestimmten Wissensgemeinschaft zuzuordnen.

Abgrenzung Schritte zum Erreichen dieses optionalen Ziels können allenfalls dann eingeleitet werden, wenn die ersten beiden Ziele erreicht worden sind und genügend Zeit dafür vorhanden sein sollte. Arbeiten in dieser Richtung sind jedoch nicht vorgesehen.

2 Projektorganisation

2.1 Projekt-Webseite

Die Webseite dieser Arbeit kann unter [6] eingesehen werden. Die Projektdaten befinden sich auf einem Speichermedium mit physikalischer Datenspiegelung. Ausserdem wird in regelmässigen Abständen ein Backup sämtlicher Daten gemacht.

2.2 Personelles

2.2.1 Projektteam

Name	Email Adresse	Telefonnummer
Mark Schmid	i99schmi@hta-be.bfh.ch	01 274 71 78

2.2.2 Betreuer

Name	Email Adresse	Telefonnummer
Dr. Jean-Pierre Caillot	pierre.caillot@hta-be.bfh.ch	031 33 55 256
Hoang-Van Chau	chau@hta-be.bfh.ch	031 33 55 257

2.2.3 Experte

Name	Email Adresse	Telefonnummer
Dr. Federico Flückiger	federico.flueckiger@bluewin.ch	091 610 85 80

2.3 Arbeitszeiten

2.3.1 Phase Reguläres Semester

Für die Arbeiten am Projekt stehen während des regulären Semesters stehen pro Woche 16 Lektionen à 45min zur Verfügung, was 12 Stunden entspricht.

Acht Stunden dieser Zeit fallen auf den Freitag, die restlichen vier sind verteilt auf weitere Wochentage.

2.3.2 Diplomphase

Während der Phase, die auf das 8.Semester folgt, kann die wöchentliche Arbeitszeit nach Bedarf erhöht werden.

2.4 Sitzungen

Bis und mit 5. Juli werden die Sitzungen zwischen dem Projektteam und den Betreuern nach Möglichkeit jede zweite Woche abgehalten. Dabei ist als Termin jeweils der Freitag von 11.00 bis 12.00 Uhr und als Ort die HTA-BE vorgesehen. Anschliessend an diese Phase werden Sitzungen einmal pro Monat veranschlagt.

Häufigkeit, Dauer und Periodizität der Sitzungen können im Rahmen des Projekts nach Bedarf jederzeit in Absprache zwischen Projektteam und Betreuer angepasst werden.

Reviews von Meilensteinen werden separat veranschlagt oder nach Bedarf auch in die vorgesehenen Sitzungen integriert.

3 Vorgehen

3.1 Zeitplanung

3.1.1 Projektphasen

Phase	Wichtigste Dokumente	Dauer (Wochen)	Datum	Meilensteine
Planung	Projektplan	2	KW 18 - 20 (02.05. - 16.05.)	
Recherche		4	KW 20 - 24 (16.05. - 13.06.)	M1, M2
Prototyping	Laborbericht	8	KW 24 - 32 (13.06. - 08.08.)	M3
Design / Spezifikation	Systemdesign, Realisierungs-konzept	6	KW 32 - 36 (08.08. - 19.09.)	M4
Implementation		6	KW 36 - 44 (19.09. - 31.10.)	M5
Testing / Bugfixing	Testplan	2	KW 44 - 46 (31.10. - 14.11.)	M6
Abschluss-arbeiten	Projektbericht	2	KW 46 - 48 (14.11. - 28.11.)	M7, M8
Reserve		4	KW 48 - 02 (28.11. - 09.01.)	
Abgabe			KW 02	M9
Total		34		

Bemerkung: Die drei Phasen *Design / Spezifikation*, *Implementation* und *Testing / Bugfixing* werden einen iterativeren Charakter besitzen, als dies aus der Tabelle hervorgeht. Die zeitlichen Rahmenbedingungen insbesondere für die Fälligkeit der Meilensteine sind demnach als Richtwerte aufzufassen.

3.1.2 Meilensteine

Meilenstein	Beschreibung
M1	Projektplan erstellt
M2	Pflichtenheft abgesegnet
M3	Prototyp erstellt
M4	Applikationsdesign abgeschlossen
M5	Abschluss der Implementation
M6	Resultate aus Testphase eingeflossen
M7	Projektbericht erstellt
M8	Gesamtpaket bereinigt und abgabebereit
M9	Abgabe

3.2 Risiken

3.2.1 Personeller Ausfall (Krankheit, Beruf, ...)

Die Eintretenswahrscheinlichkeit eines längeren personellen Ausfalls ist als gering einzuschätzen. Sollte dieser Fall trotzdem eintreten, wird aber ein grosser Einfluss auf den Zeitplan des Projekts erwartet, insbesondere aufgrund der Tatsache, dass das Projektteam personell nur aus einer Person besteht.

3.2.2 Thematisches Neuland

Die Theorie der Neuronalen Netze ist eng mit der Mathematik verknüpft, wobei durchaus unvorhergesehene Hürden auftreten können, welche eines vertieften Studiums der verwendeten mathematischen Prinzipien bedürfen. Der damit verbundene zustzliche Zeitverbrauch ist kann ein Risiko darstellen.

3.2.3 Technische Risiken

Technische Risiken wie beispielsweise Ausfall von Datenträgern sind als gering einzuschätzen. Unter 2.1 werden Massnahmen genannt, um diesem Risiko entgegenzutreten.

4 Resultate

4.1 Besonderheiten dieser Diplomarbeit

Die Thematik dieser Diplomarbeit ist nicht in erster Linie im Bereich der klassischen Software-Enticklung anzusiedeln. Es handelt sich vielmehr um eine Arbeit mit starkem experimentellem Charakter. Aus diesem Grund wird der Protokollierung von Labor- und anderen Resultaten sowie den Testinstallationen eine grosse Bedeutung zugemessen.

4.2 Dokumente

Wie bereits unter 3.1.1 erwähnt, gibt es einige für die Arbeit zentrale Dokumente:

Bezeichnung	Inhalt
Pflichtenheft	Beschreibung von Aufgabenstellung, Projektorganisation und Definition der genauen Ziele.
Realisierungskonzept	Detaillierte Beschreibung des abgestrebten Designs der Applikation
Laborbericht	Resultate der verschiedenen Experimente aus der Labor- bzw. Forschungsphase
Implementation und Testbericht	Tests, Resultate, Beurteilung
Laborjournal	Chronologische Auflistung der im Rahmen des Projektverlaufs angefallenen Tätigkeiten
Projektbericht	Abschliessender Bericht mit Themen wie Projekt-verlauf, Zielen, Analysen, Resultaten

5 Beurteilungskriterien

5.1 Bewertungsliste

Die folgende Liste basiert auf der offiziellen Bewertungsliste [4] der HTA-BE. Aufgrund des experimentellen Charakters dieser Arbeit wurde jedoch ein Punkt ergänzt und mit dem entsprechenden Gewicht versehen.

	Arbeitsschritt	Gewicht
Vorberitungsphase	Aufbau und Vollständigkeit des Pflichtenheftes	3
Durchführung	Arbeits- und Zeitplanung	3
	Kreativität, Initiative, Selbständigkeit	1
	Wahl und Anwendung der (Entwurfs-)Methodik	4
	Implementation, Robustheit, Programmierstil	2
	Laborarbeit	4
	Systemtest (Verfahren, Durchführung, Bericht)	3
	Kommunikation mit dem Experten und Betreuer	1
Ergebnis	Übereinstimmung Endprodukt/Pflichtenheft	5
	Allgemeiner Eindruck aus der Besichtigung	1
Projektbericht	Inhalt korrekt, vollständig, verständlich	3
	Sprache, Stil, Übersichtlichkeit	1
	Klare, aussagekräftige Zusammenfassung	1

Literatur

[1] http://www.hta-be.bfh.ch/∼wwwinfo/di/03/sprachenerkennungssystem.shtml

[2] http://www.languageidentifier.com/?source=overture

[3] http://www.wolfram.com/products/mathematica/index.html

[4] http://www.mathworks.com/products/matlab/

[5] http://java.sun.com/

[6] http://www.wildcard.ch/sprachen/

[7] http://www.hta-be.bfh.ch/∼wwwinfo/di/beurteilung/beurt.php3