Studienarbeit

Ein auf Flexionsformen basierendes Sprachmodell für deutsche Spracherkennung

Michael Klein
Hagsfelder Allee 18
76131 Karlsruhe

Betreuer: Ivica Rogina

Karlsruhe, den 12. Oktober 2000

Zusammenfassung

Ein wesentlicher Bestandteil der maschinellen Spracherkennung sind Sprachmodelle, die beschreiben, mit welcher Wahrscheinlichkeit bestimmte Wörter aufeinander folgen. Im Deutschen wie auch in anderen stark flektierten Sprachen ist dies jedoch aufgrund der großen Zahl verschiedener Flexionsformen schwierig. Um diesem Problem zu begegnen, wird in diesem Papier der Ansatz gemacht, jedes Wort in seine Grundform und Flektierung zu zerlegen und hieraus zwei getrennte Modelle zu erstellen, die kleiner sind und gemeinsam die Sprache besser beschreiben. In dieser Arbeit werden die dazu nötigen mathematischen Grundlagen sowie wichtige Aspekte der Implementierung ausführlich dargestellt. Abschließend wird die Qualität der so erzeugten Sprachmodelle im Spracherkenner untersucht.

Erklärung

Ich erkläre hiermit, dass ich diese Arbeit eigenhändig, ohne unzulässige Hilfsmittel und unter Angaben aller Quellen angefertigt habe.

Karlsruhe, den 12. Oktober 2000

Michael Klein

Inhaltsverzeichnis

1 Sprachmodelle für die deutsche Sprache

Ziel jedes Spracherkenners ist es, zu einer gegebenen akustischen Beoachtung X die Folge von Wörtern W' zu finden, die am wahrscheinlichsten zu dieser Beobachtung gehört. Der Erkenner maximiert also die Wahrscheinlichkeit $P(W|X)$ über alle möglichen Wortsequenzen W. Nach BAYES lässt sich diese bedingte Wahrscheinlichkeit ausdrücken als

$$\begin{aligned} W' &= \arg\max_W P(W|X) \\ &= \arg\max_W \frac{p(X|W) \cdot P(W)}{p(X)} = \arg\max_W p(X|W) \cdot P(W) \end{aligned}$$

In dieser Fundamentalgleichung der Spracherkennung lassen sich gut die zwei Hauptbestandteile eines Erkenners ausmachen: $p(X|W)$ ist das akustische Modell, welches Auskunft darüber gibt, wie wahrscheinlich es ist, X zu beobachten, wenn W gesprochen wird. Auf der anderen Seite wird $P(W)$ als Sprachmodell bezeichnet, das jedem Wort W eine Wahrscheinlichkeit zuordnet, mit der es ausgesprochen wird. Es ist selbstverständlich, dass die Erkennungsleistung eines Programms ganz wesentlich von diesen beiden Modellen abhängt. In dieser Arbeit soll auf die Erstellung von Sprachmodellen für die deutsche Sprache eingegangen werden: Zuerst werden Stardardverfahren zur Erzeugung von Sprachmodellen erläutert und Probleme aufgezeigt, die bei stark flektierten Sprachen wie dem Deutschen auftreten können. Anschließend soll ein Verfahren vorgestellt werden, welches durch Trennung von Grundform und Grammatik versucht, diesem Problem zu begegnen.

1.1 Standardansatz bei der Erstellung von Sprachmodellen

Die Wahrscheinlichkeit, eine Wortfolge $W = w_1 w_2 w_3 \dots w_n$ zu beobachten, lässt sich berechnen zu

$$\begin{aligned} P(W) &= P(w_1) \cdot P(w_2|w_1) \\ &\quad \cdot P(w_3|w_1 w_2) \cdots P(w_n|w_1 w_2 \dots w_{n-1}) \cdot P(|W| = n) \end{aligned}$$

wobei $P(w \mid W)$ die Wahrscheinlichkeit ist, das Wort w zu sehen, wenn vorher die Wortfolge W – der Kontext – gesehen wurde. Dass dieser Kontext nicht zu vernachlässigen ist, soll ein Beispiel verdeutlichen: $P(x|$ `sehr geehrter`$)$ ist für $x =$ `die` wohl sehr gering, obwohl `die` eines der häufigsten Wörter der deutschen Sprache ist. In diesem Fall würde wohl $x =$ `Herr` eine weitaus höhere Wahrscheinlichkeit liefern.
In der Praxis ist man aufgrund mangelnder Speicherkapazitäten (bisher) nicht in der Lage, zu allen möglichen Kontexten Wahrscheinlichkeiten zu allen Wörtern zu speichern. Man beschränkt sich daher auf Kontexte der Größe 1 oder 2 und

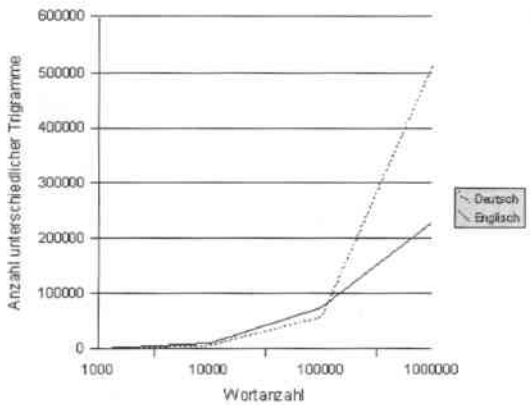

Abbildung 1: Die Anzahl unterschiedlicher Trigramme in deutschen und englischen Texten unterschiedlicher Länge. Mit wachsender Textlänge enthält Deutsch als stark flektierte Sprache erheblich mehr unterschiedliche Trigramme als Englisch.

schätzt die Wahrscheinlichkeit $P(W)$ (hier für einen Kontext von 2 Wörtern) folgendermaßen

$$P(W) \approx \prod_{i=1}^{n} P(w_i|w_{i-2}w_{i-1})$$

Diese Abschätzung ist dann plausibel, wenn man annimmt, dass Worte, die weit in der Vergangenheit gesagt wurden, sich quasi nicht mehr auf die Wahrscheinlichkeit des aktuellen Worts auswirken.
Die in der Gleichung auftretenden Wahrscheinlichkeiten $P(w_i|w_{i-2}w_{i-1})$ nennt man *Trigrammwahrscheinlichkeiten.* Diese können mit heutigen Rechnern (mit einigen Tricks) verwaltet werden, auch wenn die Zahl der theoretischen Dreierkombinationen schon bei relativ kleinen Vokabularien zu explodieren scheint.
Um tatsächlich Zahlenwerte für diese Trigrammwahrscheinlichkeiten zu erhalten, bedient man sich heute im Normalfall statistischer Methoden, das heißt man schätzt sie nach dem Gesetz der großen Zahl aus sehr großen Trainingstexten mit folgendem Bruch ab:

$$P(w_3|w_1w_2) \approx \frac{\text{Anzahl der Vorkommen von } (w_1\, w_2\, w_3) \text{ im Text}}{\text{Anzahl der Vorkommen von } (w_1\, w_2) \text{ im Text}}$$

1.2 Probleme bei stark flektierten Sprachen

Es gibt jedoch ein wesentliches Problem bei statistischen Verfahren: Viele der Dreierkombinationen werden sogar in sehr großen Trainingstexten nie oder nur einmal gesehen. Typischerweise treten 20 Prozent der gesehenen Trigramme nur ein einziges Mal auf. Welche Wahrscheinlichkeit ordnet man solchen Trigrammen zu?

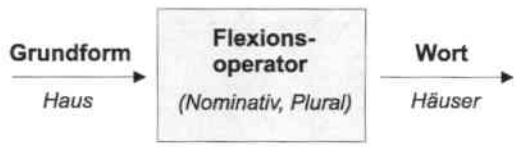

Abbildung 2: Veranschaulichung des Zusammenhangs zwischen Wort, Grundform und Flexionsoperator

Die Ursache dieses Problems liegt in der gewaltigen Zahl der sinnvollen Dreierkombinationen, die möglich sind. Bei einer Vokabulargröße von 60.000 Wörtern können theoretisch ca. $2 \cdot 10^{14}$ Trigramme auftreten, wovon ein größerer Teil als sinnvoll angesehen werden kann. Das Problem bei stark flektierten Sprachen wie Deutsch oder Koreanisch im Vergleich zu Englisch liegt nun darin, dass es aufgrund der Vielzahl der Beugungen eines Wortes sehr viel mehr sinnvolle Trigramme gibt. Ein Beispiel: Dem englischen Trigramm `(a nice house)` stehen im Deutschen gleich mehrere Trigramme - je nach Kasus - gegenüber: `(ein schönes Haus)`, `(eines schönen Hauses)`, `(einem schönen Haus)`. Weitere Beispiele ließen sich an Hand der veschiedenen Deklinationen von Verben und der Vielzahl von Pronomen (`mein, meine, meiner, meinen, meinem`...) angeben. Um diese Vermutung zu untermauern, wurde jeweils in einem deutschen und einem englischen Text mit je 1.000, 10.000, 100.000 und 1.000.000 Wörtern gezählt, wieviele unterschiedliche Trigramme auftreten (siehe Abbildung 1). Wie erwartet ist die Anzahl bei kleinen Texten für Deutsch und Englisch in etwa gleich. Bei größer werdenden Textlängen enthält das Deutsche jedoch sehr viel mehr sinnvolle Trigramme als das Englische. Daher decken deutsche Trainingstexte oft nur einen Bruchteil der möglichen grammatikalischen Formen ab und verteilen die Wahrscheinlichkeitsmasse relativ 'ungerecht' auf die zufällig gesehenen.

1.3 Ein verbesserter Ansatz: Trennung von Grundform und Flektierung

Ein Ansatzpunkt, um diesem Problem zu begegnen besteht darin, ein Wort nicht als eine atomare Einheit zu sehen, sondern es in zwei Teile zu zerlegen: seine grammatikalische Grundform und die Flektierung. Diese Flektierung kann als Operator verstanden werden, der – angewendet auf die Grundform – eben das Wort zurückliefert (siehe Abbildung 2). Im Folgenden wird daher von Grundform und Flexionsoperator gesprochen.

Eine wichtige Voraussetzung wird sein, dass jedes Wort eindeutig durch eine Grundform und einen Flexionsoperator bestimmt ist – eine Voraussetzung, die von wenigen Ausnahmen abgesehen, erfüllt ist. Diese wichtige Eigenschaft bie-

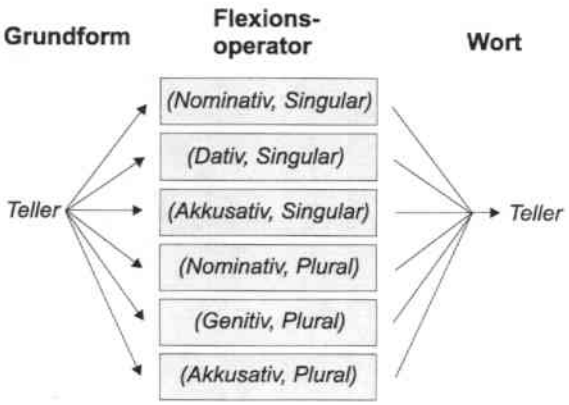

Abbildung 3: Das Wort `Teller` kann durch Anwendung von 6 verschiedenen Flexionsoperatoren aus der Grundform `Teller` entstehen.

tet nun die Möglichkeit, Trigrammwahrscheinlichkeiten nicht auf den Wörtern selbst, sondern getrennt voneinander für Grundformen einerseits und Flexionsoperatoren andererseits zu bestimmen. Der Vorteil ist offensichtlich: Im Deutschen wie auch in anderen stark flektierten Sprachen ist die Zahl der Grundformen weitaus geringer als die Zahl der auftretenden Wörter und auch die Zahl der unterschiedlichen Flexionsoperatoren wird im Vergleich hierzu eher klein sein. In der Implementierung werden daher genau zwei Trigrammmodelle verwendet: das Grundform-Modell und das Flexionsoperator-Modell, die unabhängig voneinander aus einem Trainingstext erstellt werden. Um später wieder auf Wahrscheinlichkeiten von flektierten Wörtern zu kommen, werden beide Modelle getrennt voneinander befragt und daraus eine Gesamtwahrscheinlichkeit berechnet. Hierzu wird angenommen, dass die Grundform eines Wortes kaum von den Flexionen der Vorgänger abhängt. So modelliert das Grundformmodell vor allem die Semantik und das Flexionsoperatormodell vor allem die Syntax. Ein solcher Ansatz zur Verschmelzung zweier oder mehrerer Sprachmodelle ist nicht grundsätzlich neu, sondern wird auch in anderen Bereichen eingesetzt. Es ist zu erwarten, dass hierdurch die Perplexität reduziert wird.

Um diese beiden Modelle geeignet trainieren zu können, besteht ein Hauptbestandteil der Arbeit darin, ein beliebiges Wort in seine Grundform und den Flexionsoperator zu zerlegen. In einigen Fällen mag dies einfach sein, wie zum Beispiel beim Wort `ist`, welchem eindeutig die Grundform `sein` und der Flexionsoperator `(3. Person Singular Präsens Indikativ)` zugeordnet werden kann. In den meisten Fällen ist die Grundform zwar noch eindeutig zu bestimmen, doch könnte das Wort durch Anwendung verschiedener Flexionsoperatoren hierauf gebildet worden sein (siehe Abbildung 3). Prinzipiell ist es sicher nicht

weiter schwierig, auch hieraus ein Flexionsoperatormodell zu berechnen, jedoch wird ein solches Modell schnell unscharf, da in jedem Trigramm zu viele Operatoren in Betracht gezogen werden. Schöner wäre es, man könnte jedem Wort relativ sicher einen einzigen (höchstens einige wenige) Operatoren zuordnen, zumal für das Training der Modelle grammatikalisch korrekte Texte verwenden werden, in denen jedes einzelne Wort in seinem Kontext eine eindeutige grammatikalische Form darstellt. Einen Großteil der Arbeit macht daher ein spezieller Parser aus, der jedes Wort mit Wörtern seiner Umgebung vergleicht, um möglichst genau die gerade vorliegende Flexion zu bestimmen. Dieser Vorgang wird in 3.2 genauer erläutert.

1.4 Vorteile bei der Datennutzung

Ein großer Vorteil dieses Ansatzes wird die bessere Nutzung der Trainingsdaten sein. Bei einem gewöhnlichen Trigrammmodell erhöht ein gesehenes Trigramm $(w_1\ w_2\ w_3)$ nur die Wahrscheinlichkeit für dieses Trigramm. Trennt man jedoch Grundform und Flexionsoperator auf und trainiert getrennte Modelle, so lassen sich gesehene grammatikalische Zusammenhänge auch auf andere Grundformen übertragen - insgesamt werden also durch ein gesehenes Trigramm gleich mehrere Wahrscheinlichkeiten auch anderer Trigramme mitverbessert. Auf der anderen Seite werden die Trigrammwahrscheinlichkeiten für Grundformen viel genauer sein, da sich die Trainingsdaten auf viel weniger unterschiedliche Grundformen konzentrieren können und nicht von grammatikalischen Formen 'abgelenkt' werden.

2 Mathematische Fundierung des Ansatzes

Im folgenden Abschnitt werden die mathematischen Grundlagen des Verfahrens genauer erläutert und die wichtigsten Formeln hergeleitet. Zuerst wird die dazu nötige Nomenklatur in 2.1 vorgestellt. Welche Unabhängigkeitsannahmen für die Herleitung nötig sind, wird dann in 2.2.1 bis 2.2.3 aufgelistet und jeweils mit statistischen Untersuchungen untermauert. Die eigentliche Herleitung folgt dann in 2.3.

2.1 Klassen- und Überklasseneinteilung der Wörter

Wie schon gesehen, lässt sich zu jedem Wort w des Vokabulars V eine Grundform $gf(w)$ angeben. Hiermit lässt sich also eine Klasseneinteilung auf V definieren:

$$c_w := \{x \in V | gf(x) = gf(w)\}$$

Die Klasse eines Wortes beinhaltet also neben dem Wort selbst alle weiteren Flektierungen dieses Wortes. Die Klasse des Wortes `nehmen` würde also wie

folgt aussehen:

$$c_{nehmen} = \{nehmen, nehme, nimmst, nimmt, nahm, genommen, \ldots\}$$

Diese Klasseneinteilung ist nur dann sinnvoll, wenn wir Folgendes voraussetzen:

Jedes Wort aus V gehört zu genau einer Klasse.

Diese Annahme ist zwar weitestgehend erfüllt, es gibt jedoch einige Wörter, deren Grundform sich ohne Kontext nicht eindeutig bestimmen lässt. Es gilt zum Beispiel `weiß` $\in c_{weiß}$ und `weiß` $\in c_{wissen}$. Zur Vereinfachung wird in solchen Fällen das Wort beliebig, aber verbindlich einer Klasse zugeordnet.
Da alle Wörter einer Wortart ähnlich geformte Klassen bilden, kann man sie zu Überklassen, den Wortartklassen $\mathcal{W}$ zusammenfassen. Es gilt also zum Beispiel

$$\mathcal{W}_{Substantiv} = \{c_{Haus}, c_{Baum}, c_{Spracherkenner}, \ldots\}$$

$$\mathcal{W}_{Verb} = \{c_{nehmen}, c_{erkennen}, c_{sein}, \ldots\}$$

Da jede Überklasse durch einen Repräsentanten eindeutig bestimmt ist, wird für sie im Folgenden auch die Schreibweise $\mathcal{W}(Repraesentant)$ verwendet. Es gilt also zum Beispiel

$$\mathcal{W}(c_{Haus}) = \mathcal{W}(c_{Baum}) = \mathcal{W}_{Substantiv}$$

Jeder Wortart (also Überklasse) $\mathcal{W}$ steht ein charakteristischer Satz von Flexionoperatoren $\mathcal{F}^{\mathcal{W}} = \{f_1^{\mathcal{W}}, f_2^{\mathcal{W}}, \ldots\}$ zur Verfügung, der nur auf Klassen $c \in \mathcal{W}$ anwendbar ist. Jeder dieser Operatoren bildet eine Wortklasse $c \in \mathcal{W}$ auf ein Wort $w \in c$ ab, er 'flektiert' also die Grundform c. Formal gilt:

$$\begin{aligned} f_i^{\mathcal{W}} &: c_w \mapsto w' \\ \text{wobei } &c_w \in \mathcal{W} \\ \text{und } &w' \in c_w \end{aligned}$$

Mit Hilfe dieses Operatorensatzes lässt sich jede Klasse $c_w \in \mathcal{W}$ beschreiben. Es gilt:

$$c_w = \bigcup_{f \in \mathcal{F}^{\mathcal{W}}} f(c_w)$$

Analog zum Operator gf von oben, der zu einem gegebenen Wort dessen Grundform zurückliefert, kann man jetzt den Operator $flek$ definieren, der zu einem gegebenen Wort w die Menge F der Operatoren bestimmt, die - angewendet auf die Grundform des Wortes $gf(w)$ - gerade das Wort w selbst wieder ergeben:

$$flek : w \mapsto F := \{f_i | f_i(c_w) = w\} \subseteq \mathcal{F}^{\mathcal{W}(c_w)}$$

2.2 Mathematische Herleitung zur Berechnung der Trigrammwahrscheinlichkeiten

Mit dieser Beschreibung lässt sich nun die Formel für die Trigrammwahrscheinlichkeiten herleiten: Das Sprachmodell soll - wie jedes Standardtrigrammmodell auch - die Wahrscheinlichkeit

$$P(w_3|w_1\ w_2)$$

für drei beliebige Wörter $w_1, w_2, w_3 \in V$ vorhersagen. Diese Wahrscheinlichkeit lässt sich zerlegen in zwei Teile: Wie wahrscheinlich ist es, die Klasse von w_3 zu beobachten, nachdem w_1 und w_2 beobachtet wurden und wie wahrscheinlich ist es dann, aus dieser Klasse tatsächlich das Wort w_3 auszuwählen, auch unter der Voraussetzung, dass zuvor w_1 und w_2 gesehen wurden. Formal gilt also

$$P(w_3|w_1\ w_2) = P(c_{w_3}|w_1\ w_2) \cdot P(w_3|c_{w_3},\ w_1\ w_2) \tag{1}$$

Um diese Formel vereinfachen zu können, müssen einige Unabhängigkeitsannahmen getroffen werden. Diese sind im ersten Moment rein willkürlich, weshalb versucht wird, sie durch statistische Untersuchungen plausibel zu machen.

2.2.1 Bedingte Wahrscheinlichkeit einer Klasse

Für den ersten Teil der Formel nehmen wir Folgendes an:

> Unabhängigkeitsannahme 1
> Die Wahrscheinlichkeit einer Wortklasse hängt nur von den Klassen der Vorgängerworte ab, nicht jedoch von deren Flektierung.

Oder formal:

$$P(c_{w_3}|w_1\ w_2) = P(c_{w_3}|c_{w_1}\ c_{w_2}) \tag{2}$$

Diese Annahme ist deshalb sinnvoll, da die Klasse eines Wortes im Wesentlichen den Inhalt des Wortes repräsentiert und von den jeweiligen Flektierungen abstrahiert. Ein Beispiel:

```
der amerikanische Präsident
den amerikanischen Präsidenten
der deutsche Bundeskanzler
des deutschen Bundeskanzlers
```

Offensichtlich sind jeweils die ersten beiden und die letzten beiden Trigramme inhaltlich sehr ähnlich, obwohl diese jeweils sechs verschiedene Worte verwenden. Ist man nun zum Beispiel an der Wahrscheinlichkeit für die Klasse $c_{Bundeskanzler}$ interessiert, wenn vorher w_1 und w_2 beobachtet wurden, so reicht

	MESSWERT	ABS. ABWEICHUNG
$P(c_{der}\|c_{sei}\ c_{in})$	**0.577**	-
$P(c_{der}\|ist\ in)$	0.549	0.028
$P(c_{der}\|waren\ in)$	0.616	0.039
$P(c_{der}\|sind\ in)$	0.649	0.072
$P(c_{der}\|seien\ in)$	0.522	0.055
$P(c_{der}\|sei\ in)$	0.557	0.020
$P(c_{neu}\|c_{in}\ c_{der})$	**0.0077**	-
$P(c_{neu}\|in\ der)$	0.0017	0.0060
$P(c_{neu}\|in\ die)$	0.0021	0.0056
$P(c_{neu}\|in\ das)$	0.0167	0.0150
$P(c_{neu}\|in\ dem)$	0.0035	0.0042

Tabelle 1: Begründung von Unabhängigkeitsannahme 1. Zur Bestimmung der Wahrscheinlichkeiten wurden 2 Mio. Wörter aus deutschen Nachrichtentexten verwendet. Man stellt fest, dass der Unterschied zwischen $P(c_{w_3}|w_1\ w_2)$ und $P(c_{w_3}|c_{w_1}\ c_{w_2})$ bis auf wenige Ausreißer gering ist.

es sicher aus, sich nur auf die Klasseninformation von w_1 und w_2 zu verlassen, denn diese enthalten ja gerade die inhaltliche Information. Wir bestimmen also $P(c_{Bundeskanzler}|c_{w_1}\ c_{w_2})$ und erwarten für $c_{w_2} = c_{amerikanisch}$ einen niedrigen und $c_{w_2} = c_{deutsch}$ einen hohen Wert.
Auch Untersuchungen an Texten zeigen, dass diese Unabhängigkeitsannahme weitestgehend plausibel ist. Für die Ergebnisse in Tabelle 1 wurden Nachrichtentexte deutscher Zeitungen zu einem Textkörper mit 2 Millionen Wörtern zusammengefügt (größere Textkörper mit z.B. 100 Mio. Wörtern wurden auf Grund knapper Ressourcen nicht verwendet) und daraus die entsprechenden Wahrscheinlichkeiten mittels

$$P(z|x\ y) = \frac{\#(x\ y\ z) \text{ im Text}}{\#(x\ y) \text{ im Text}}$$

abgeschätzt. Es zeigt sich, dass sich $P(c_{w_3}|w_1\ w_2)$ und $P(c_{w_3}|c_{w_1}\ c_{w_2})$ tatsächlich nur wenig unterscheiden und somit in der Praxis für eine Berechnung von $P(w_3|w_1\ w_2)$ im Prinzip austauschbar sind.

2.2.2 Verteilung der Elemente in einer Klasse

Für den zweiten Teil der Wahrscheinlichkeitsformel (1) $P(w_3|c_{w_3},\ w_1\ w_2)$ lassen sich zwei Unabhängigkeiten angeben. Die erste bezieht sich auf $P(w_3|c_{w_3})$, also die Wahrscheinlichkeit, ein Element aus einer gegebenen Klasse herauszuziehen. Wir nehmen an, dass diese Wahrscheinlichkeit nicht von der konkreten Grundform abhängt, sondern für alle Klassen dieser Wortart gleich ist.

> Unabhängigkeitsannahme 2
> Die Wahrscheinlichkeit für einen Flexionsoperator bei gegebener Grundform hängt nur von der Wortart dieser Grundform ab.

Formal soll also für alle $w \in V$ gelten

$$P(w \mid c_w) = P(flek(w) \mid \mathcal{W}(c_w)) \tag{3}$$

Dass diese Annahme weitestgehend erfüllt ist, sollen wieder statistische Untersuchungen an großen Texten belegen (siehe Tabelle 2). Wir betrachten exemplarisch die Wortart der Verben ($\mathcal{W}_{Verb}$) und greifen hieraus die häufigsten acht Grundformen heraus. Alle diese Grundformen flektieren wir mit dem gleichen Satz von Flexionsoperatoren und bestimmen von den entstehenden Formen die Wahrscheinlichkeiten mit Hilfe des großen Textkörpers aus 2.2.1. Es ergibt sich, dass Worte, die aus gleichen Flexionsoperatoren entstanden sind, auch ähnliche Wahrscheinlichkeiten aufweisen. Nur das Wort `können` hat eine etwas andere Verteilung. Besonders gut stimmen die Modalverben `haben` und `sein` überein, die in deutschsprachigen Texten einen Großteil der Verben stellen.

2.2.3 Bedingte Wahrscheinlichkeit einer Flektierung

Die letzte Unabhängigkeitsannahme, die getroffen werden soll, ist das Gegenstück zu Annahme 1 und spiegelt die Unabhängigkeit der grammatikalischen Flexion vom Inhalt wider. Sie lautet

> Unabhängigkeitsannahme 3
> Die Wahrscheinlichkeit des Flexionsoperators eines Wortes hängt nur von den Flexionsoperatoren der Vorgängerworte ab, nicht jedoch von deren Grundformen.

Das bedeutet formal

$$P(flek(w_3) \mid w_1\ w_2) = P(flek(w_3) \mid flek(w_1)\ flek(w_2))$$

Einsichtig wird diese Annahme durch ein kleines Beispiel:

```
die grüne Tasse
eine zerbrochene Tasse
die gefüllte Tasse
ein rotes Tasse
die zerbrochenen Tasse
dem gefüllten Tasse
```

Flexions-Operator	Grundform			
	sein	haben	sollen	können
1.Pers. sg. Präs.	bin/ist	habe/hat	soll	kann
3.Pers. sg. Präs.	0.4643	0.5082	0.4533	0.3476
2.Pers. sg. Präs.	bist	hast	sollst	kannst
	0.0000	0.0002	0.0000	0.0000
1.Pers. pl. Präs.	sind	haben	sollen	können
3.Pers. pl. Präs.	0.2686	0.2471	0.3249	0.4003
1.Pers. sg. Prät.	war	hatte	sollte	konnte
3.Pers. sg. Prät.	0.1708	0.1736	0.1477	0.1342
2.Pers. sg. Prät.	warst	hattest	solltest	konntest
	0.0000	0.0000	0.0000	0.0000
1.Pers. pl. Prät.	waren	hatten	sollten	konnten
3.Pers. pl. Prät.	0.0961	0.0733	0.0741	0.1172

Flexions-Operator	Grundform			
	wollen	bleiben	stellen	geben
1.Pers. sg. Präs.	will	bleibe/bleibt	stelle/stellt	gebe/gibt
3.Pers. sg. Präs.	0.5427	0.4250	0.3108	0.4891
2.Pers. sg. Präs.	willst	bleibst	stellst	gibst
	0.0000	0.0000	0.0000	0.0000
1.Pers. pl. Präs.	wollen	bleiben	stellen	geben
3.Pers. pl. Präs.	0.3257	0.3589	0.3677	0.2145
1.Pers. sg. Prät.	wollte	blieb	stellte	gab
3.Pers. sg. Prät.	0.0632	0.1226	0.2323	0.2597
2.Pers. sg. Prät.	wolltest	bliebst	stelltest	gabst
	0.0000	0.0000	0.0000	0.0000
1.Pers. pl. Prät.	wollten	blieben	stellten	gaben
3.Pers. pl. Prät.	0.0685	0.0935	0.0892	0.0363

Tabelle 2: Zu Unabhängigkeitsannahme 2. Beispielhaft wurden für die Wortart $\mathcal{W}_{Verb}$ die acht häufigsten Grundformen aufgegriffen und jeweils die Wahrscheinlichkeiten für verschiedene Flexionen an Hand eines Trainingstextes abgeschätzt. Man beobachtet, dass Flexionen, die aus gleichen Flexionsoperatoren entstanden sind, ähnliche Wahrscheinlichkeitswerte aufweisen, unabhängig von der vorliegenden Grundform. Insbesondere die Modalverben sein und haben weisen große Ähnlichkeit auf.

Die ersten drei Trigramme sind grammatikalisch korrekt, während die letzten drei grammatikalisch falsch sind. Dies kann jedoch schon allein aus den Flexionsoperatoren der beiden Kontextwörter abgelesen werden, der textuelle Inhalt, also ob die Tasse grün, zerbrochen oder gefüllt ist, spielt hierfür keine Rolle. Dies lässt sich gut erkennen, wenn man die (korrekten und gekürzten) Flexionsoperatoren von jedem Wort bestimmt. Obiges Beispiel sieht dann so aus:

```
(ART,bestimmt,fem,sg,Nom) (ADJ,fem,sg,Nom) (SUB,fem,sg,Nom)
(ART,unbestimmt,fem,sg,Nom) (ADJ,fem,sg,Nom) (SUB,fem,sg,Nom)
(ART,bestimmt,fem,sg,Nom) (ADJ,fem,sg,Nom) (SUB,fem,sg,Nom)
(ART,unbestimmt,neut,sg,Nom) (ADJ,neut,sg,Nom) (SUB,fem,sg,Nom)
(ART,bestimmt,fem,pl,Nom) (ADJ,fem,pl,Nom) (SUB,fem,sg,Nom)
(ART,bestimmt,mask,sg,Dat) (ADJ,mask,sg,Dat) (SUB,fem,sg,Nom)
```

Hierbei steht `ART` für Artikel, `ADJ` für Adjektiv und `SUB` für Substantiv; `mask`, `fem` und `neut` geben den Genus, `sg` und `pl` den Numerus, `Nom` und `Dat` den Kasus des Wortes an. Beim Training des Modells erwarten wir, dass die ersten drei Trigramme – da aus korrekten Textstücken entstanden – häufig vorkommen, während die letzen drei selten oder gar nicht auftreten und somit geringe Wahrscheinlichkeiten für die drei grammatikalisch nicht korrekten Trigramme von oben bewirken.

2.3 Zusammensetzung und Diskussion der Grundformel

Mit diesen Unabhängigkeitsannahmen lässt sich jetzt die Formel zur Berechnung der Trigrammwahrscheinlichkeiten $P(w_3|w_1\ w_2)$ zusammensetzen. Betrachten wir dazu nochmal Formel (1)

$$P(w_3|w_1\ w_2) = P(c_{w_3}|w_1\ w_2) \cdot P(w_3|c_{w_3},\ w_1\ w_2)$$

Den ersten Teil ersetzen wir nach Unabhängigkeitsannahme 1 und erhalten

$$P(w_3|w_1\ w_2) = P(c_{w_3}|c_{w_1}\ c_{w_2}) \cdot P(w_3|c_{w_3},\ w_1\ w_2)$$

Nach Unabhängigkeitsannahme 2 und 3 ergibt sich dann für den zweiten Teil

$$P(w_3|w_1\ w_2) = P(c_{w_3}|c_{w_1}\ c_{w_2}) \cdot P(flek(w_3)\,|\,\mathcal{W}(c_{w_3}),\ flek(w_1)\ flek(w_2)) \quad (4)$$

Diese Form legt den Weg nahe, wie eine solche Trennung von Grundform und Grammatik realisiert werden sollte. Im Prinzip werden zwei getrennte Modelle verwendet: Das Grundform-Modell, welches auf Trigrammen aus Grundformen trainiert wird und damit Wahrscheinlichkeiten für den ersten Teil der Formel liefert und das Flexionsoperatoren-Modell, welches auf Trigrammen aus Flexionsoperatoren trainiert wird und – nach Wortarten getrennt – die Wahrscheinlichkeiten für den zweiten Teil der Formel liefert. Die Auftrennung nach Wortarten ist deshalb nötig, da nicht jeder Flektionsoperator auf jede Wortart anwendbar ist (z.B. existiert kein Akkusativ von Verben).

2.4 Notwendige Erweiterungen aufgrund von Ambiguitäten

Ein Problem existiert in Formel (4) noch durch den Operator *flek*. Dieser liefert zu einem Wort w gleich eine ganze Menge F von Flexionsoperatoren, welche das Wort w aus seiner Grundform c_w bilden würden. Wie bestimmt man jedoch $P(\mathcal{C}|\mathcal{A}\,\mathcal{B})$, wenn

$$\begin{aligned}\mathcal{A} &= \{a_1, a_2, \ldots, a_l\},\\ \mathcal{B} &= \{b_1, b_2, \ldots, b_m\} \text{ und}\\ \mathcal{C} &= \{c_1, c_2, \ldots, c_n\}\end{aligned}$$

endliche Mengen sind und das Modell nur Wahrscheinlichkeiten von Einzeltrigrammen zur Verfügung stellt?
Die Menge $\mathcal{C}$ stellt kein Problem dar, sie resultiert in eine Summe von Wahrscheinlichkeiten:

$$P(\mathcal{C}|\mathcal{A}\,\mathcal{B}) = \sum_{k=1}^{n} P(c_k|\mathcal{A}\,\mathcal{B})$$

Für die Mengen $\mathcal{A}$ und $\mathcal{B}$ müssen alle möglichen Zweierkombinationen $(a\ b)$ von Elementen aus diesen Mengen betrachtet werden und die zugehörigen Wahrscheinlichkeiten $P(c|a\ b)$ gewichtet aufsummiert werden. Als Gewicht dient jeweils die Wahrscheinlichkeit, diese Kombination aus allen Kombinationen auszuwählen. Es gilt also:

$$P(c|\mathcal{A}\,\mathcal{B}) = \sum_{i=1}^{l}\sum_{j=1}^{m} \left(P(c|a_i\ b_j) \cdot P(a_i\ b_j|\mathcal{A}\,\mathcal{B}) \right)$$

Insgesamt ergibt sich:

$$P(\mathcal{C}|\mathcal{A}\,\mathcal{B}) = \sum_{k=1}^{n}\sum_{i=1}^{l}\sum_{j=1}^{m} \left(P(c_k|a_i\ b_j) \cdot P(a_i\ b_j|\mathcal{A}\,\mathcal{B}) \right)$$

Mit dieser Rechenvorschrift lassen sich auch die Wahrscheinlichkeiten für den zweiten Teil der Grundformel (4) bestimmen.

3 Umsetzung im Programm

In diesem Abschnitt wird erklärt, wie die vorgestellten Konzepte in ein lauffähiges Programm übertragen worden sind. Hierzu werden nur die Grundstrukturen und Hauptprobleme deutlich gemacht, auf Implementationsfeinheiten oder Programmcode wird verzichtet. Das Programm ist vollständig in C++ verfasst und rein objektorientiert programmiert. Es besteht im wesentlichen aus vier Komponenten: Wort, welches eine Grundform mit allen Flexionen repräsentiert (siehe 3.1), Lexikon, das alle Wörter verwaltet, Tafel, ein Parser, der ein gegebenes Wort durch Kontextanalyse in Grundform und Flexionsoperator zerlegt (siehe 3.2) und Modell, das die getrennten Trigrammmodelle bereitstellt und die angefragten Wahrscheinlichkeiten berechnet (siehe 3.3).

3.1 Wort - Repräsentation des Wortschatzes

3.1.1 Struktur von Wort

Die Oberklasse zur Verwaltung aller deutschen Grundformen ist die Klasse Wort, welche alle Wortarten als Unterklassen besitzt. Diese repräsentieren jeweils eine Grundform der entsprechenden Wortart und stellen alle zugehörigen Flexionen zur Verfügung. Die verwendeten Wortarten sind Substantiv, Verb, Adjektiv, Artikel, Präposition, allg.Pronomen, Personalpronomen, Reflexivpronomen, Possessivpronomen und Unflektiert, welche alle unflektierten Wortarten wie Numerale, Konjunktionen, Interjektionen, Adverben und unbekannte Wörter aufnimmt. Die Unterscheidung nach Wortarten ist notwendig, da für Worte einer Wortart charakteristische Attribute verwaltet werden müssen. Für ein Verb sind beispielsweise Person und Tempus wichtig, während für ein Adjektiv Eigenschaften wie Kasus und Deklinationstyp entscheidend sind. Welche Attribute im einzelnen für jede Wortart aufgenommen wurden, soll im nächsten Abschnitt erläutert werden.

3.1.2 Betrachtete Attribute

Um Wörter aus beliebigen Wortarten miteinander vergleichen zu können, beispielsweise um deren Flektierung im Kontext genau bestimmen zu können (siehe auch 3.2), wurden vier *Kernattribute* für alle verwendeten Wortarten festgelegt. Diese sind Person, Numerus, Kasus und Genus. Sie stehen für jedes Wort zur Verfügung oder werden sinnvoll festgelegt. Hierzu zwei Beispiele: Ein Verb hat an sich keinen Kasus, bezieht sich jedoch immer auf das Subjekt, welches im Nominativ steht. Für Verben wird daher der Kasus auf Nominativ festgesetzt. Bei Substantiven fehlt rein grammatikalisch das Attribut Person, sie werden jedoch im Satz immer mit Verben in der 3. Person verbunden. Die 3. Person ist daher für Substantive festgelegt.
Jede Wortart hat nun noch *spezielle Attribute*, die alleine für diese Wortart Sinn

WORTART	SPEZIELLE ATTRIBUTE	TYP
Adjektiv	Deklinationstyp	flexionsbezogen
	Steigerung	flexionsbezogen
Verb	Tempus	flexionsbezogen
	Modus	flexionsbezogen
	Partizip mit haben/sein	grundformbezogen
Artikel	Bestimmtheit	grundformbezogen
Possessivpronomen	Besitzgenus	flexionsbezogen
	Besitznumerus	flexionsbezogen

Tabelle 3: Spezielle Atrribute von Wortarten. Die Bestimmtheit bei Artikel kann auch als flexionsbezogen angesehen werden, wenn nur eine einzige Artikel-Grundform vorliegt, die in bestimmte und unbestimmte Artikel flektiert werden kann.

machen. Diese sind in Tabelle 3 zusammengefasst. Allgemein können Attribute in zwei Klassen unterteilt werden: flexionsbezogene und grundformbezogene Attibute. Flexionsbezogene Attribute können für verschiedene Flexionen einer Grundform auch unterschiedliche Werte annehmen. Grundformbezogene Attibute haben dagegen den gleichen Wert für alle Flexionen einer Grundform. Ein Beispiel für ein flexionsbezogenes Attribut ist Genus bei Adjektiven. Die Flexion `schöner` hat das Geschlecht maskulin, während `schönes` das Geschlecht neutrum besitzt. Im Lexikon müssen daher für jede Grundform die Flexionen aller Kombinationen von Werten flexionsbezogener Attribute zur Verfügung gestellt werden. Ein grundformbezogenes Attribute ist Genus bei Substantiven. Unabhängig von der Flexion `Haus`, `Häuser`, `Hauses`, das Geschlecht ist bei `Haus` immer Neutrum, was bereits an der Grundform erkennbar ist. Es ist offensichtlich, dass bei gegebenem Wort die Bestimmung flexionsbezogener Attribute bedeutend schwieriger ist, als die grundformbezogener Attribute.

3.2 Tafel - Zerlegung in Grundform und Flektierung

Nach Grundformel (4) müssen zwei unterschiedliche Modelle aus einem Trainingstext erstellt werden: das Grundwort- und das Flexionsoperatoren-Modell. Es ist daher nötig, jedes Wort des Trainigstextes in seine Grundform und seine Flexionsoperatoren (mit den entsprechenden Attributen aus 3.1.2) zu zerlegen. Im einfachsten Fall schlägt man dazu einfach in einem grammatikalischen Wörterbuch nach (siehe 3.2.1) und beschafft sich so die Grundform und mögliche Operatoren. Oft kann das Wörterbuch jedoch nicht eindeutig bestimmen, welcher Flexionsoperator genau vorliegt - ein Abgleich des Wortes mit seinen benachbarten Worten im Satz (dem *Kontext*) wird nötig. Dies erledigt die Tafel. Ihre Funktionsweise wird in 3.2.2 und 3.2.3 genau erläutert. In 3.2.4 werden an Hand von Beispieltexten der einfache Wörterbuchnachschlag und die Verwendung der Tafel miteinander verglichen. Warum die Eindeutigkeit der Zerlegung

für die Erstellung eines Sprachmodells so wichtig ist, wird in 3.2.5 erörtert.

3.2.1 Einfacher Wörterbuchnachschlag

Die grammatikalischen Informationen eines Wortes sind im *Grammatikwörterbuch* gespeichert. Für diese Implementierung wurde das *Bielefelder Lexikon* verwendet, das im Rahmen des VERBMOBIL-Projekts entwickelt wurde [5]. In ihm sind zu jedem Wort dessen Grundform und alle in Frage kommenden Flexionsoperatoren gespeichert. Nach der Annahme aus 2.1 gehört jedes Wort zu genau einer Grundform und damit zu genau einer Wortart. Daher kann das Wörterbuch hier immer eine eindeutige Auskunft geben. Schwieriger wird es für den Flexionsoperator: Hier kann das Wörterbuch nur eine Menge von möglichen Operatoren zurückliefern, ohne angeben zu können, welcher der Operatoren in diesem Fall tatsächlich vorliegt. Das Wort `die` hat vier mögliche Flexionsoperatoren, jenachdem ob es sich auf etwas Weibliches im Singular oder etwas Männliches im Plural bezieht. Auch der Kasus ist noch nicht eindeutig bestimmbar. Solche Wörter sind nicht die Ausnahme - im Gegenteil: Es gibt nur sehr wenige flektierte Worte, die allein durch Nachschlagen im Wörterbuch eindeutig bestimmbar sind. Beispiele hierfür sind die Wörter `bin`, `bist` und `ist`, bei denen Person, Numerus, Modus und Tempus schon ohne weitere Analyse eindeutig sind.

3.2.2 Genauere Bestimmung der Flektierung durch Kontextanalyse auf der Tafel

Um den vorliegenden Flexionsoperator genauer bestimmen zu können, ist man auf den Kontext angewiesen. Steht zum Beispiel das Wort `die` im Zusammenhang von `die Frau`, so kann der Genus eindeutig als feminin und der Numerus auf Singular angegeben werden. Für den Kasus stehen nur noch zwei Möglichkeiten offen: Nominativ oder Akkusativ. Auch die Analyse des Wortes `Frau` profitiert hiervon: als Kasus können der Genitiv und der Dativ ausgeschlossen werden, sonst müsste es ja `der Frau` heißen. Im Normalfall gleicht man also die Flexionsoperatoren eines Wortes mit denen der Wörter in der Nachbarschaft ab, um zu einem gemeinsamen genaueren Satz von Operatoren zu gelangen. Hilfreich hierfür sind dabei offensichtlich solche Wörter, die schon durch einen reinen Wörterbuchnachschlag relativ genau bestimmt werden können. Die Hauptfrage, die sich hierbei stellt, ist jedoch: Welche Wörter gehören zusammen, etwa weil sie das Subjekt oder ein Objekt des Satzes bilden? Wo fängt ein Präpositionalgefüge an, wo hört ein Nebensatz auf? Besondere Probleme bereiten außerdem Verben: Sie stehen mit dem Subjekt des Satzes in Beziehung, müssen jedoch nicht unbedingt im Satz in dessen Nähe stehen. Ein Standardlösungsansatz hierzu wäre, eine formale Grammatik der deutschen Sprache aufzustellen und jeden Satz in einem Parser in seine Bestandteile zerlegen zu lassen. Dies scheitert jedoch an zwei Problemen: Zum einen

ist die deutsche Sprache sehr komplex. Im Gegensatz zum Englischen können die Satzglieder praktisch in beliebiger Reihenfolge zu Sätzen zusammengesetzt werden. Auch eingeschobene Nebensätze und unvollständige Sätze sind keine Seltenheit. Eine formale Grammatik hierfür wäre sehr umfangreich und deckte doch nicht alle Sonderfälle ab. Zum anderen soll das Verfahren mit großen Wortschätzen, etwa Zeitungstexten umgehen können. Da alle Vokabularien (besonders auch bei einem Einsatz im Spracherkenner) beschränkt sind, treten in Trainingstexten häufig Wörter auf, die nicht im Vokabular enthalten sind, sogenannte *out-of-vocabulary words (OOVs)*. Diese können die Analyse des Parsers zunichte machen.

Für diesen Ansatz soll ein anderes Verfahren zum Einsatz kommen: die Tafel. Sie arbeitet nach einem einfachen Prinzip: Die Wörter des Trainingstextes werden der Reihe nach auf die Tafel geschrieben. Diese entscheidet an Hand einer Menge von Heuristiken (siehe 3.2.3), welche Wörter wahrscheinlich zusammengehören und sammelt diese in einem gemeinsamen Behälter. Um die Entscheidung (besonders für die Verbzuordnung) zu erleichtern, ist die Zahl der unterschiedlichen Behälter relativ klein. Die Implementierung arbeitet mit 5 bis 10 Behältern. Soll ein Wort auf die Tafel geschrieben werden, welches in keinen der vorhandenen Behälter eingefügt werden kann, so blockiert sie. Um wieder Platz zu schaffen, müssen die zuerst auf die Tafel geschriebenen Worte wieder gelesen werden. Für sie kann eine relativ genaue Flexionsoperatormenge angeben werden: Sie berechnet sich als Schnitt aller am Behälter beteiligten Flexionsoperatormengen. Verglichen werden können solche Flexionsoperatoren an Hand der Kernattribute (siehe 3.1.2), die auch für Operatoren unterschiedlicher Wortarten gleich sind.

Dieser Ansatz hat mehrere Vorteile: Zum einen wirken sich OOVs nicht so drastisch auf das Bestimmungsergebnis aus, da die Tafel nicht auf jedes einzelne Wort angewiesen ist und daher solche OOV-Worte einfach überspringen kann. Auch die Probleme mit verschachtelten Sätzen und unterschiedlichen Objektreihenfolgen im Satz sind für die Tafel nicht direkt sichtbar. Weiterhin kann die Tafel mit ihrem einfachen Konzept relativ effizient große Datenmengen verarbeiten.

3.2.3 Verwendete Heuristiken

Bei der Verteilung der Wörter geht die Tafel nach folgendem Grundprinzip vor: Aufeinander folgende Wörter werden so lange im selben Behälter abgelegt, bis der Schnitt aller Flexionsoperatoren leer ist, also die Wörter rein grammatikalisch nicht mehr syntaktisch zusammengehören können. Das letzte Wort kommt dann in den nächsten freien Behälter. Für eine Reihe von Wörtern würde ein Abgleichen auf der Tafel keinen Vorteil bedeuten. Hierzu gehören insbesondere generell unflektierte Wortarten (Konjunktionen, Interjektionen, Numerale,

Adverbiale), aber auch unflektierte Adjektivformen. Sie werden daher nicht in einem Behälter der Tafel gespeichert.

Gewisse Wörter können nun die Tafel von ihrem Grundprinzip abweichen und schon vorzeitig einen neuen Behälter beginnen lassen. Hierzu gehören vor allen Dingen Konjunktionen, die mehrere Satzteile oder ganze Sätze miteinander verbinden. Sie sind daher ein sicheres Indiz dafür, dass ein neues Satzglied beginnt. Auch Präpositionen künden im Regelfall den Beginn einer Präpositionalphrase an, die in einem neuen Behälter abgelegt werden sollte.
Oft sind in Trainingstexten die Satzgrenzen, welche ähnlich wie Konjunktionen neue Satzglieder anzeigen, nicht durch spezielle Symbole (Punkt, Komma, Semikolon etc.) gekennzeichnet. Trifft man daher auf ein groß geschriebenes Wort, das normalerweise nicht groß geschrieben wird, so ist dies ein Hinweis auf einen Satzanfang. Ein neuer Behälter sollte begonnen werden.
Hilfreich für die Behälterzuordnung sind auch Wortartpaare, die niemals zusammen zu einem Satzglied gehören können. Hierzu gehören Pronomen und Substantive (Pronomen stehen ja gerade für ein Nomen), Pronomen und Artikel sowie zwei Artikel. Außerdem gibt es Wortarten die stets alleine in einem Behälter stehen, wie Personalprononem und Reflexivpronomen.

Verben spielen eine gewisse Sonderrolle bei der Zuordnung auf der Tafel. Sie sind grammatikalisch gesehen ein eigenes Satzglied, sind aber eng mit dem Subjekt des Satzes verknüpft. Oftmals ist daher das Verb die einzige Möglichkeit, den Kasus verschiedener Objekte genau zu bestimmen, andererseits wird die Person des Verbs oft erst durch einen Vergleich mit dem Subjekt klar. Problematisch ist dabei, dass Subjekt und Verb nicht zwangsläufig in räumlicher Nähe zueinander im Satz stehen müssen. Gerade Nebensätze beginnen oft mit dem Subjekt, während das Verb ganz am Ende steht. Aus diesem Grund werden Verben stets in einem speziellen Behälter (zum Beispiel Behälter 0) gespeichert[1]. Erst wenn ein weiteres Verb auf die Tafel geschrieben werden soll oder vermutlich ein neuer Satz beginnt (z.B. wegen einer Konjunktion oder eines groß geschriebenen Wortes, siehe oben), wird das Verb an das Objekt in einem der regulären Behälter gebunden, welches am wahrscheinlichsten dem Subjekt des Satzes entspricht. Behälter, die kein Objekt enthalten, sondern zum Beispiel eine Präpositionalphrase und Behälter, die ein Objekt enthalten, welches sicher kein Nominativ ist, können dabei übersprungen werden. Zu allen anderen wird mit Hilfe eines Maßes eine Ähnlichkeit aufgrund der übereinstimmenden Kernattribute berechnet. Behälter, die nur ein Personalpronomen beinhalten, erhalten wegen ihrer besonderen Häufigkeit als Subjekt hierbei einen Bonus. An den Behälter mit der größten Ähnlichkeit wird anschließend das Verb gebunden. Die Abbildungen 4 und 5 veranschaulichen die Funktionsweise der Tafel noch einmal ausführlich an einem Beispiel.

[1] Verbformen wie Partizipien und Befehlsformen werden nicht auf der Tafel gespeichert

Abbildung 4: Beispiel für die Funktionsweise der Tafel. Der Satz "`Bei der Verteilung der Wörter geht die Tafel nach folgendem Grundprinzip vor`" soll auf die Tafel geschrieben werden. In jedem Schritt sind dabei die Belegung der Behälter und die zugehörigen Schnitte der Flexionsoperatoren in drei der vier Kernattribute in der Form (Kasus, Genus, Numerus) angegeben. [1] Die groß geschriebene Präposition `Bei` signalisiert einen Satzanfang. Sie wird daher in einen leeren Behälter gelegt, hier in 1. Behälter 0 ist ausschließlich für Verben reserviert. Mögliche Flexionsoperatoren für `bei` sind (akk,*,*) und (dat,*,*). [2] Der Artikel `der` hat die Flexionsoperatoren (nom,mask,sg), (dat,fem,sg), (gen,fem,sg), (gen,*,pl), bildet also mit den Flexionsoperatoren aus Behälter 1 keinen leeren Schnitt. Er wird demnach hier abgelegt. Als neuer Schnitt ergibt sich (dat,fem,sg). Hierzu passt auch `Verteilung` mit den Flexionsoperatoren (*,fem,sg). [3] Der Artikel `der` wird im nächsten Behälter gespeichert, da nie zwei Artikel in einem Behälter stehen. Das Wort `Wörter` mit den Flexionsoperatoren (nom,neut,pl), (gen,neut,pl), (akk,neut,pl) ergibt den nichtleeren Schnitt (gen,neut,pl) und wird deshalb ebenfalls in Behälter 2 abgelegt. [4] Das Verb `geht` wird im ausgezeichneten Behälter 0 abgelegt. Seine Flexionsoperatoren sind (nom,*,sg). Weiter in Abbildung 5.

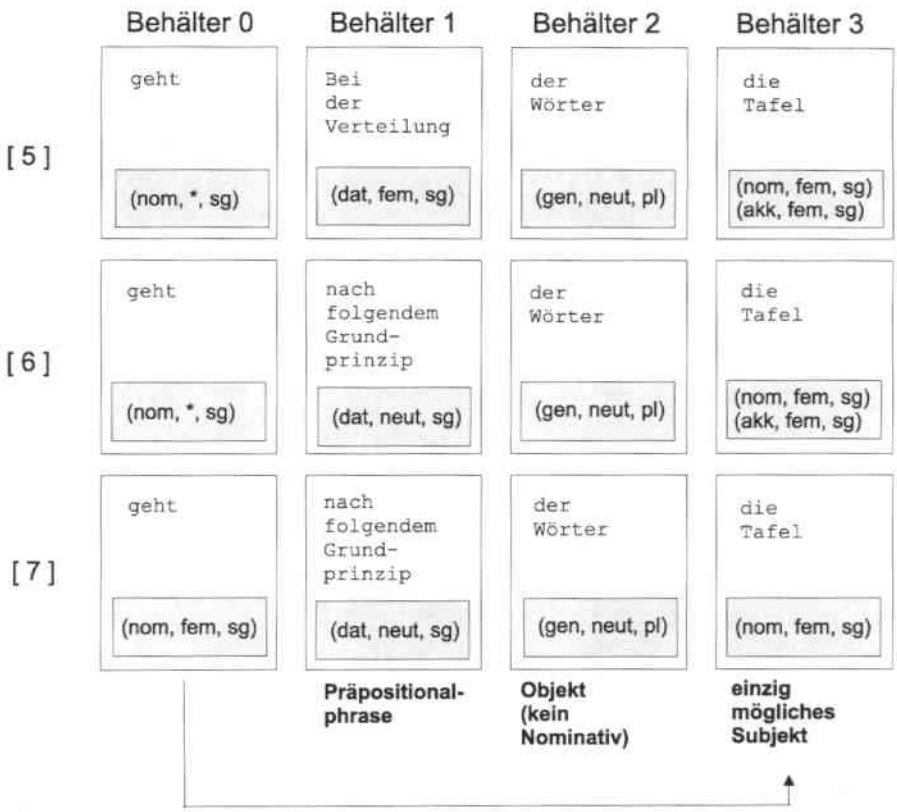

Abbildung 5: Fortsetzung des Beispiels. [5] `die` und `Tafel` belegen den neuen Behälter 3. [6] `nach` benötigt als Präposition einen leeren Behälter – die Wörter `Bei`, `der` und `Verteilung` werden von der Tafel genommen. Ihre Flexionsoperatoren sind jetzt in den Kernattributen eindeutig bestimmt: (dat, fem, sg). In den leeren Behälter 1 werden `nach`, `folgendem` und `Grundprinzip` eingefügt. Die Präposition `vor` wird als unflektiertes Wort nicht auf die Tafel geschrieben. [7] Der Satz ist beendet, das zum Verb in Behälter 0 gehörige Subjekt wird gesucht. Da Behälter 1 eine Präpositionalphrase und Behälter 2 ein Genitivobjekt enthält, kommt nur Behälter 3 in Frage. Als Flexionsoperator für Verb und Subjekt ergibt sich damit (nom,fem,sg). Anschließend können alle Wörter von der Tafel genommen werden.

	WÖRTERBUCH	TAFEL
UNK	91	91
unflektiert	63	63
korrekt, eindeutig	26	105
korrekt, 2 Operatoren	27	23
korrekt, 3 Operatoren	48	24
korrekt, 4 Operatoren	26	14
korrekt, 5 Operatoren	1	3
korrekt, 6 Operatoren	37	7
korrekt, 7 Operatoren	0	0
korrekt, 8 Operatoren	14	1
korrekt, 9 Operatoren	7	0
korrekt, 10-19 Op.	11	4
korrekt, 20-29 Op.	1	0
korrekt, $\geq$ 30 Op.	10	0
durchschn. Op.-Zahl	**6.63**	**2.16**
falsch (Kasus)	0	26
falsch (Genus)	0	1
falsch (Numerus)	0	1
falsche Wortart	4	4

Tabelle 4: Bestimmung von Flexionsoperatoren auf einem Nachrichtentext mit 367 Wörtern einerseits mit einfachem Wörterbuchnachschlag, andererseits zusätzlich mit Kontextanalyse auf der Tafel. Mehr als viermal so viele Wörter werden hierbei korrekt und eindeutig bestimmt.

3.2.4 Vergleich zwischen Wörterbuchnachschlag und Tafel

Um die Leistungsfähigkeit der Tafel zu testen, wurde versucht, zu 367 Wörtern aus einem Nachrichtentext die jeweiligen Flexionsoperatoren zu bestimmen. Auf größere Texte wurde wegen der aufwändigen Handmarkierung verzichtet. Einerseits wurde nur im Wörterbuch nachgeschlagen, andererseits wurde eine Kontextanalyse mit Hilfe der Tafel durchgefhrt. Die Ergebnisse sind in Tabelle 4 und Abbildung 6 präsentiert. Man erkennt, dass viermal mehr Wörter korrekt und eindeutig bestimmt werden und sehr große Unsicherheiten von acht und mehr Flexionsoperatoren nur noch sehr selten vorkommen. Andererseits treten im Gegensatz zum Wörterbuchnachschlag Fehlbestimmungen der Flexionsoperatoren auf, die in mehr als 90 Prozent der Fälle Fehler im Kasus sind.

3.2.5 Vorteil der eindeutigen Bestimmung

Es stellt sich die Frage, wieso eine möglichst eindeutige Bestimmung der Flexionsoperatoren tatsächlich eine Verbesserung im daraus entstehenden Sprach-

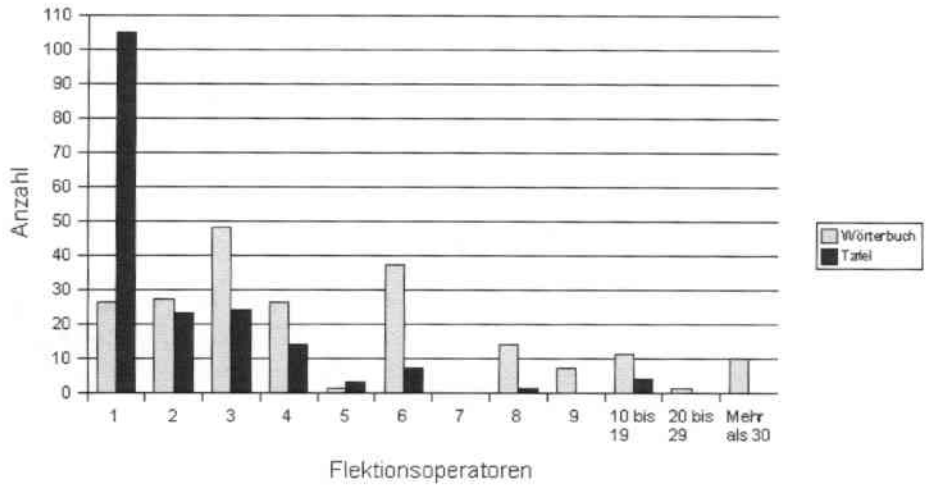

Abbildung 6: Grafische Darstellung der Ergebnisse aus Tabelle 4. Aufgetragen sind nur die korrekt bestimmten Flexionsoperatoren.

modell bringt. Nehmen wir drei Wörter, deren Flexionsoperatoren wir nicht eindeutig bestimmen können. In das Modell müssen dann alle Kombinationen von Flexionsoperatortrigrammen aufgenommen werden. Dies sind ohne Kontextanalyse der Tafel im Schnitt nach Tabelle 4 $6.6^3 \approx 287$ Trigramme, von denen nur eins korrekt ist. Alle anderen repräsentieren in gewisser Weise falsche grammatikalische Zusammenhänge, die später auf andere Grundformen übertragen werden. Ein solches Modell zeichnet sich durch eine hohe Entropie aus, da die Wahrscheinlichkeitsmasse relativ gleichmäßig auf korrekte und falsche Trigramme verteilt wird. Mit Hilfe der Kontextanalyse sinkt die Entropie drastisch: Im Schnitt werden pro Worttripel nur ca. $2.16^3 \approx 10$ Trigramme aufgenommen. Die Wahrscheinlichkeitsmasse konzentriert sich jetzt viel stärker auf die wenigen korrekten Dreiertupel.

3.3 Modell - Bereitstellung der Wahrscheinlichkeiten

3.3.1 Interne Repräsentation durch zwei Standardmodelle

Um Wahrscheinlichkeiten nach Grundformel (4) berechnen zu können, müssen die zwei Teilwahrscheinlichkeiten

$$P(c_{w_3}|c_{w_1}\ c_{w_2})$$

und

$$P(flek(w_3)\,|\,\mathcal{W}(c_{w_3}),\ flek(w_1)\ flek(w_2))$$

bereitsgestellt werden. Hierzu dient die Klasse Modell, welche einerseits als Grundform-, andererseits als Flexionsoperatormodell instanziiert wird. Um diese Modelle trainieren zu können, werden große deutschsprachige Texte mit Hilfe von Lexikon und Tafel in Grundform und Flexionsoperator zerlegt und die dabei auftretenden Häufigkeiten von Tri-, Bi- und Unigrammen im jeweiligen Modell gespeichert. Bei genügend großer Anzahl kann daraus später mittels

$$P(c_{w_3} \mid c_{w_1}\, c_{w_2}) = \frac{\#(c_{w_1}\, c_{w_2}\, c_{w_3}) \text{ im Text}}{\#(c_{w_1}\, c_{w_2}) \text{ im Text}}$$

die Wahrscheinlichkeiten des Grundformmodells abschätzen.

Problematischer ist das Flexionsoperatorenmodell. Die Bestimmung der Flexionsoperatoren ist – selbst bei Verwendung der Tafel – nicht immer eindeutig. Stehen für die Wörter w_i $(i = 1, 2, 3)$ jeweils n_i mögliche Flexionsoperatoren zur Verfügung, so ergeben sich daraus insgesamt $n_1 \cdot n_2 \cdot n_3$ mögliche Trigramme, von denen in der Regel jedoch nur eines das Richtige ist. Um dieses korrekte Dreiertupel nicht auszulassen, werden alle $n_1 \cdot n_2 \cdot n_3$ Trigrame in das Modell aufgenommen, wobei ihre Häufigkeit jedoch mit dem Faktor $\frac{1}{n_1 \cdot n_2 \cdot n_3}$ gewichtet wird. Somit gehen eindeutig bestimmte Trigramme sehr stark, weniger gut bestimmte entsprechend abgeschwächt in das Flexionsoperatormodell ein. Für jedes Trigramm aus eindeutigen Flexionsoperatoren steht somit eine gewichtete Häufigkeit zur Verfügung, mit deren Hilfe man die Wahrscheinlichkeiten nach 2.4 folgendermaßen berechnen kann. Seien

$$\begin{aligned} \mathcal{C} &= flek(w_3) = \{c_1, c_2, \ldots, c_n\} \\ \mathcal{A} &= flek(w_1) = \{a_1, a_2, \ldots, a_l\} \\ \mathcal{B} &= flek(w_2) = \{b_1, b_2, \ldots, b_m\} \end{aligned}$$

Dann gilt

$$\begin{aligned} P(\mathcal{C} \mid \mathcal{W}(c_{w_3}),\, \mathcal{A}\,\mathcal{B}) &= \sum_{k=1}^{n} \sum_{i=1}^{l} \sum_{j=1}^{m} P(c_k \mid \mathcal{W}(c_{w_3}),\, a_i\, b_j) \cdot P(a_i\, b_j | \mathcal{A}\,\mathcal{B}) \\ &= \sum_{k=1}^{n} \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{\#(a_i\, b_j\, c_k)}{\#(a_i\, b_j\, \mathcal{W}(c_{w_3}))} \cdot \frac{\#(a_i\, b_j)}{\sum_{s=1}^{l} \sum_{t=1}^{m} \#(a_s\, b_t)} \end{aligned}$$

Neben den gewöhnlichen Trigrammen aus drei eindeutigen Flexionsoperatoren müssen also auch Häufigkeiten zu Dreiertupeln der Form (Flexionsoperator, Flexionsoperator, Wortart) gespeichert werden, die angeben, wie häufig eine bestimmte Wortart $\mathcal{W}(c_{w_3})$ auftrat, wenn zuvor die Flexionsoperatoren a_i und b_j gesehen wurden. Das Flexionsoperatormodell ist daher im allgemeinen größer als das Grundformmodell.

3.3.2 Verbesserungen am Modell

Wie bei anderen Sprachmodellen auch treten beim Grundform- und Flexionsoperatormodell Probleme durch zu wenige Trainingsdaten auf. Viele der möglichen Trigramme treten gar nicht oder nur ein einziges Mal auf, was es schwierig macht, für sie eine fundierte Wahrscheinlichkeit nach

$$P(c \mid a\, b) = \frac{\#(a\, b\, c)}{\#(a\, b)}$$

zu schätzen (*Zero Frequency Problem*). Ein Ansatz, der auch im Fall mehrerer Modelle zum Erfolg führt, ist die Glättung der Daten. Ziel der Glättung ist es, die gegebenen Wahrscheinlichkeiten so anzupassen, dass sie der Wirklichkeit genauer entsprechen. Eine Vorgehensweise, um dies zu erreichen, ist die Einbeziehung von Bi-, Uni- und Zerogrammen, um die Wahrscheinlichkeiten nicht ausschließlich aus Häufigkeiten von Trigrammen abschätzen zu müssen. Für die Implementierung wurde die JELINEK-MERCER-Glättung [1] verwendet, die Tri-, Bi-, Uni- und Zerogrammwahrscheinlichkeiten mit Hilfe eines Mischungsfaktors λ $(0 < \lambda < 1)$ interpoliert. Es gilt

$$\begin{aligned} P_{JM}(c \mid a\, b) = \quad & \lambda P(c \mid a\, b) + (1 - \lambda) \\ \cdot \quad & \{\lambda P(c \mid b) + (1 - \lambda) \\ \cdot \quad & (\lambda P(c) + (1 - \lambda) \cdot P_{zero})\} \end{aligned}$$

wobei P_{zero} die Zerogrammwahrscheinlichkeit ist, die zu

$$P_{zero} = \frac{1}{|V|}$$

mit $|V|$ als Vokabulargröße berechnet werden kann. Diese Zerogrammwahrscheinlichkeit sorgt unter anderem dafür, dass keine Trigrammwahrscheinlichkeit 0 werden kann, denn selbst wenn weder Tri-, Bi- noch Unigramm im Trainingstext aufgetreten sind, gilt

$$P_{JM}(c|a\, b) \geq (1 - \lambda)^3 \cdot P_{zero} = (1 - \lambda)^3 \cdot \frac{1}{|V|} > 0$$

Um den Parameter λ geeignet zu bestimmen, können mehrere Verfahren eingesetzt werden. Die bekanntesten Vertreter sind der BAUM-WELCH-Algorithmus und "deleted interpolation" [1].

3.3.3 Probleme bei der Verwendung des Modells im Spracherkenner

Beim Einsatz eines flexionsbasierten Sprachmodells in einem Erkenner tritt ein Problem auf: Die Berechnung der Wahrscheinlichkeiten nach Grundformel (4) ist sehr zeitaufwändig. Jedes Wort muss mit Hilfe des Grammatikwörterbuches

in Grundform und Flektionsoperator zerlegt werden, danach wird das Grundformmodell dreimal, das Flektionsoperatormodell aufgrund von Ambiguitäten sogar öfter konsultiert. Abschließend werden alle Teilergebnisse nach Bestimmung von Gewichtungsfaktoren verrechnet. Da Trigrammwahrscheinlichkeiten sehr häufig vom Erkenner angefordert werden, ist es unbedingt erforderlich, dass diese sehr schnell vorliegen. Die jetzige Implementierung kann daher nicht direkt im Erkenner zur on-the-fly Berechnung der Wahrscheinlichkeiten verwendet werden. Es ist daher nötig, das Sprachmodell an die Standards heutiger Spracherkenner anzupassen, die zur Leistungssteigerung das komplette Sprachmodell vor der Erkennung in vorberechneter Form (Wort1 Wort2 Wort3 Wahrscheinlichkeit) einlesen. Jedes Trigramm muss also explizit aufgeführt werden. Als Quasi-Standard für solche Sprachmodelldateien hat sich das ARPA-Format von DOUG PAUL durchgesetzt (Spezifikation siehe [2]), welches darüberhinaus alle Wahrscheinlichkeiten zur schnelleren Multiplikation in logarithmierter Form enthält. Trennt man nun die Modelle in ein Grundform- und ein Flexionsoperatormodell auf, so lassen sich zwar nach der Grundformel (4) alle Trigrammwahrscheinlichkeiten $P(w_3 \mid w_1\, w_2)$ berechnen, jedoch sind diese nicht explizit im Modell abgelegt. Nur Grundformtrigramme sind noch vorhanden. Will man jetzt ein solches zweigeteiltes Modell zum Beispiel im ARPA-Format ausgeben, so stellt sich die Frage, welche der möglichen Trigramme tatsächlich mit ihrer Wahrscheinlichkeit in das Modell aufgenommen werden sollen. Es bieten sich verschiedene Möglichkeiten:

1) Alle Kombinationen $(w_1\, w_2\, w_3)$ mit $w_i \in V$ aufnehmen
Dieser Ansatz scheitert an der gewaltigen Anzahl von Kombinationen. Bei einem durchschnittlichen Vokabular mit 65.000 Wörtern müssten schon ca. $2.75 \cdot 10^{14}$ Trigramme gespeichert werden, von denen die allermeisten sinnlos wären und nur eine Wahrscheinlichkeit nahe 0 hätten.

2) Nur Trigramme $(w_1\, w_2\, w_3)$ in das Model aufnehmen, welche im Trainingstext T tatsächlich gesehen wurden
Dieser Ansatz hat das Problem, dem verbesserten Sprachmodell nicht gerecht zu werden und seine Stärken zu vernachlässigen. Gerade für sinnvolle Trigramme, die nicht unbedingt im Text gesehen wurden, können gute Wahrscheinlichkeiten abgeschätzt werden, indem die Grammatik aus ähnlichen Fällen wiederverwendet wird. Ein Beispiel: Gesehen wurde `das kleine Haus` und `die hellen Lichter`. Auch die sinnvollen, obwohl nicht gesehenen Trigramme `die kleinen Häuser` und `das helle Licht` erhalten hohe Wahrscheinlichkeiten.

3) Die gesehenen Grundformtrigramme $(c_{w_1}\, c_{w_2}\, c_{w_3})$ in allen Flektierungen aufnehmen
Dies ist die logische Erweiterung zu Ansatz 2, hat jedoch ein ähnliches Problem wie Ansatz 1: Die Anzahl der entstehenden Trigramme wird riesig, wovon sehr viele grammatikalisch sinnlos sind.

4) Die gesehenen Grundformtrigramme $(c_{w_1}\, c_{w_2}\, c_{w_3})$ mit allen gesehenen, passenden Flexionsoperatortrigrammen flektieren und aufnehmen
Dieser Ansatz stellt einen Kompromiss zwischen Größe des entstehenden Modells und Vernachlässigung von Stärken dar. Er wird daher auch in der Implementation verwendet. Es ist jedoch zu beachten, dass ein Modell, welches in ein solches Format gepresst wurde, nicht die gleiche Leistungsfähigkeit zeigen kann wie zwei getrennte Modelle, mit deren Hilfe man effektiv über alle $|V|^3$ Trigramme verfügen kann.

4 Experimente

Im folgenden Abschnitt sollen einige nach diesem Ansatz erstellte Sprachmodelle bewertet werden. Dies geschieht zum einen in 4.1 theoretisch-formal mittels Perplexitätsmessungen, andererseits sollen die Modelle auch praktischen Tests in einem Spracherkenner unterzogen werden. Dies wird in Abschnitt 4.2 behandelt.

4.1 Theoretische Bewertung der Sprachmodelle an Hand von Perplexitäten

4.1.1 Definition Perplexität

Um die Güte eines flexionsbasierten Modells im Vergleich zu einem Standardtrigrammmodell formal bewerten zu können, ist ein Maß nötig. Sehr verbreitet ist die *Perplexität* eines Sprachmodells bezüglich eines gegebenen Testtexts, welcher ausreichend groß und repräsentativ sein sollte. Anschaulich gesprochen lässt man die Wahrscheinlichkeit bestimmen, mit der das Sprachmodell diesen Testtext als mögliche Sequenz von Wörtern ansieht. Es gilt also für den Trainingstext $T = t_1\, t_2 \ldots t_n$ und das Sprachmodell M:

$$P_M(T) = P_M(t_1) \cdot P_M(t_2 \,|\, t_1) \cdot \prod_{i=3}^{n} P_M(t_i \,|\, t_{i-1}\, t_{i-2})$$

Gute Sprachmodelle sollten auf dem repräsentativen Trainingstext T eine relativ hohe Wahrscheinlichkeit $P_M(T)$ erzielen. Hieraus errechnet sich dann die Perplexität B zu

$$B = P_M(T)^{-1/n}$$

Sie gibt an, aus wievielen Wörtern der Spracherkenner im Schnitt das korrekte Wort herausfinden muss. Man stebt also kleine Perplexitäten an.

4.1.2 Verwendete Trainingsdaten

Zur Erstellung der Sprachmodelle wurden drei unterschiedliche Trainingstexte verwendet:

Trainingstext germ100M.1

- Länge: 100.000.000 Wörter
- Quelle: deutsche Nachrichtentexte und Wettervorhersagen
- Komposita (wie z.B. `Dreikönigstreffen`) in Einzelwörter zerlegt
- Wortverbindungen (wie z.B. `SPD-Vorsitzender`) in Einzelwörter zerlegt
- Keine Satzzeichen. Satzanfang und -ende durch Spezialsymbole `<s>` und `</s>` markiert

Trainingstext germ100M.2

- wie germ100M.1, Komposita jedoch nicht aufgetrennt

Trainingstext sz1400K

- Länge: 1.400.000 Wörter
- Quelle: Nachrichtentexte der Süddeutschen Zeitung des Jahres 1995, keine Wettervorhersagen
- Komposita und Wortverbindungen nicht zerlegt. Text grammatikalisch korrekt.
- Keine Satzzeichen. Satzanfang und -ende durch Spezialsymbole `<s>` und `</s>` markiert

Alle drei Trainingstexte verwenden die alte Rechtschreibung, um mit dem verwendeten Grammatikwörterbuch übereinzustimmen. Auf diesen Texten wurden einerseits vier verschiedene Trigrammmodelle auf herkömmliche Weise mit dem CMU-Toolkit von RONI ROSENFELD berechnet [2]. Hierbei wurde ein Cutoff von 1 für Bi- und Trigramme verwendet und die Daten linear geglättet. Andererseits wurden die vier Modelle durch Auftrennung von Grundform und Flexionsoperator erzeugt. Ein Cutoff von 1 wurde hierbei jeweils für Grundform und Flexionsoperator eingestellt. Alle acht Modelle verwendeten ein Vokabular von

TRAININGSDATEN	STANDARDMODELL MIT CMU-TOOLKIT	FLEXIONSBASIERTES MODELL
die ersten 10.000 Wörter von germ100M.1	germ10K.1.reg.arpa	germ10K.1.arpa
die ersten 1 Mio. Wörter von germ100M.1	germ1M.1.reg.arpa	germ1M.1.arpa
die ersten 1 Mio. Wörter von germ100M.2	germ1M.2.reg.arpa	germ1M.2.arpa
sz1400K	sz1400K.reg.arpa	sz1400K.arpa

Tabelle 5: Die acht verwendeten Sprachmodelle im Überblick

ca. 65.000 Wörtern und wurden im ARPA-Format ausgegeben. Eine Übersicht über alle Sprachmodelle findet sich in Tabelle 5.

4.1.3 Verwendete Testdaten

Die Perplexitäten wurden auf sechs verschiedenen Testtexten berechnet. Davon waren zwei Nachrichtentexte anderer Zeitungen, die dem späteren Einsatzgebiet des Modells entsprechen, drei Texte aus anderen Genren und der Text, der dem Spracherkenner in Abschnitt 4.2 in gesprochener Form zur Erkennung vorgelegt wurde. Tabelle 6 zeigt alle Testtexte auf einen Blick.

NAME	LÄNGE	GENRE	QUELLE
rz20K	20.000 Wörter	Nachrichten	Artikel der Rhein-Zeitung (alle Rubriken) von Jan. 1996 bis März 1996
welt23K	23.000 Wörter	Nachrichten	alle Artikel der WELT vom 01.06.1996
werther31K	31.000 Wörter	Anspruchsvolle Litertur	*Die Leiden des jungen Werther* von GOETHE
schnee3K	3.000 Wörter	Einfache Literatur	Kindermärchen *Schneewittchen* der GEBRÜDER GRIMM
phil10K	10.000 Wörter	wissenschaftliche Arbeit	Diplomarbeit im Fach Philosophie
correct2K	1.545 Wörter	Text für den Erkenner	105 Nachrichtensätze deutscher Zeitungen

Tabelle 6: Die sechs Testtexte im Überblick: 2 Nachrichtentexte, 3 Texte anderer Genre und der Text, der dem Spracherkenner in gesprochener Form vorgelegt wurde.

	germ1M.1.arpa	germ1M.1.reg.arpa	germ1M.2.arpa	germ1M.2.reg.arpa
rz20K	1172.7	344.3	1187.4	342.3
welt23K	1462.8	405.5	1572.1	406.1
werther31K	4725.6	1495.1	4886.5	1424.6
schnee3K	3291.8	746.5	3362.7	724.0
phil10	3664.6	1148.1	3686.0	1111.1
correct2K	1135.6	344.3	1156.5	339.0

	sz1400K.arpa	sz1400K.reg.arpa
rz20K	1448.1	318.7
welt23K	1342.4	350.4
werther31K	2095.0	815.1
schnee3K	1982.2	655.6
phil10	2357.2	744.5
correct2K	1301.4	302.1

Tabelle 7: Gemessene Perplexitäten der Sprachmodelle.

4.1.4 Gemessene Perplexitäten

Die Perplexitäten wurden mit Hilfe des CMU-Toolkits [2] berechnet. Dabei wurden Wahrscheinlichkeiten der Form $P(<\texttt{UNK}> \mid w_1\, w_2)$ nicht mitgerechnet, da solche beim späteren Einsatz im Spracherkenner nicht angefragt werden. Die Ergebnisse sind in Tabelle 7 zusammengefasst.

Es fällt auf, dass die Perplexitäten der flexionsbasierten Modelle bedeutend höher sind als die der Standardtrigrammmodelle. Dies überrascht im ersten Moment, da doch hierfür offensichtlich mehr Information (etwa das Grammatikwörterbuch) verwendet wurde. Man kann jedoch die spätere Leistung im Erkenner nicht direkt aus der Perplexität folgern, da die akustische Verwechselbarkeit der Wörter nicht in die Berechnung einbezogen wird. Aussagekräftiger ist die sogenannte *A-Perplexität*, bei der das Vokabular auf solche Wörter eingeschränkt wird, die akustisch leicht zu verwechseln sind. Diese ist jedoch sehr schwierig zu berechnen und wird daher kaum verwendet. Es zeigt sich im nächsten Abschnitt, dass bei der tatsächlichen Verwendung im Spracherkenner, die flexionsbasierten Modelle ungefähr gleiche Fehlerraten liefern, im Falle von sz1400K.arpa sogar etwas bessere.

4.2 Einsatz der Sprachmodelle im Spracherkenner

4.2.1 Schwierigkeiten

Beim Einsatz eines Sprachmodells in einem realen Spracherkenner treten verschiedene Probleme auf. Ein zentrales Problem ist die Größe des Vokabulars,

die bei heutigen Erkennern in der Regel einen Wert von $2^{16} = 65.536$ nicht überschreiten darf, um aus Speicherplatzgründen alle Wörter eindeutig mittels eines 2-Byte-Schlüssels adressieren zu können. In der Regel ist jedoch die Anzahl aller Flexionen von nur wenigen Grundformen bedeutend größer, so dass eine Auswahl getroffen werden muss, welche Flexionen man tatsächlich in das Vokabular aufnimmt. Es bietet sich an, alle Flexionen einer Grundform entweder komplett aufzunehmen oder komplett auszulassen, um später nicht Gefahr zu laufen, Formen bilden zu müssen, die nicht Teil des Vokabulars sind.

Ein ähnliches Problem stellt das Aussprachelexikon dar. Hierin sind zu allen Wörtern mögliche Aussprachevarianten in einer Art Lautschrift vermerkt, die im Erkenner mit der 'gehörten' Phrase verglichen wird. Grundsätzlich gilt also, dass der Erkenner nur solche Wörter verstehen kann, die in seinem Aussprachelexikon enthalten sind. Da nun in den seltensten Fällen dieses Aussprachelexikon und das Grammatikwörterbuch die gleichen Wörter enthalten, müssen beide aufeinander abgestimmt werden, um ein optimales Ergebnis zu erzielen. Man geht dabei wie folgt vor: Wörter, die nur im Grammatikwörterbuch zu finden sind, werden nicht ins Vokabular aufgenommen. Auch alle ihre Flexionen entfallen. Enthält andererseits das Aussprachelexikon ein Wort, welches nicht im Grammatikwörterbuch zu finden ist, so wird dieses dem Grammatikwörterbuch als Wort der Wortart Unvollständig zugefügt und bei der Berechnung der Wahrscheinlichkeiten wie ein unflektierbares Wort behandelt. Da für solche Wörter keine grammatikalischen Eigenschaften wie zum Beispiel Genus bekannt sind, entstehen durch sie bei der Erkennung häufig Grammatikfehler (siehe 4.2.2).

4.2.2 Erkennungsleistung

Für die folgenden Experimente wurde der Spracherkenner JANUS verwendet, dem der Text correct2K (siehe auch Tabelle 6) von vier verschiedenen Sprechern vorgelesen wurde. correct2K besteht aus 105 Nachrichtensätzen, die deutschen Zeitungen entnommen sind und hat insgesamt 1.545 Wörter. Zwischen dem erkannten Text und correct2K wurde satzweise die Wortfehlerrate berechnet.

Im ersten Durchlauf wurden das flexionsbasierte Sprachmodell germ1M.1.arpa und als Vergleich dazu das Standardtrigrammmodell germ1M.1.reg.arpa verwendet. Mit germ1M.1.reg.arpa machte der Erkenner 44,2 Prozent Fehler, mit germ1.1.arpa 45,4 Prozent. Das flexionsbasierte Modell schnitt also etwas schlechter ab. Untersucht man die Art der Fehler, die durch das flexionsbasierte Modell entstanden, so erkennt man einen deutlichen Trend: Häufig werden zwei Substantive erkannt, die grammatikalisch scheinbar sinnlos aneinander gereiht sind.

Ein Beispiele soll das verdeutlichen:

korrekt:
`[...] pochen die Politiker auf eine Art Gleichbehandlung`
erkannt:
`[...] brauchen die Politiker auf einer Amtsgericht Behandlung`

Der letzte Teil des erkannten Satzes ist grammatikalisch falsch[2]. Erklären lässt sich dies durch den verwendeten Trainingstext, in dem zusammengesetzte Substantive (Komposita) und Wortverbindungen, in denen ein Bindestrich auftritt (wie zum Beispiel `SPD-Vorsitzender`) in einzelne Worte aufgetrennt wurden, um bessere Vokabularabdeckung zu erzielen. Im allgemeinen entstehen jedoch hierdurch grammatikalisch falsche Konstrukte, die das Sprachmodell dann später auf andere Sachverhalte falsch überträgt. Ursache für die Grammatik des erkannten Satzes könnte zum Beispiel folgende Wortfolge im Trainingstext sein:

gesehen:
`auf einer Schiff Schaukel`
entstanden aus:
`auf einer Schiffschaukel`

Ursprünglich war die Folge grammatikalisch mit (Präposition, Artikel-dat-fem-sg, Substantiv-dat-fem-sg) korrekt, die zerlegte Folge enthält zusätzlich noch das eigentlich unpassende sächliche Wort `Schiff`, hat also die Flexionsoperatorfolge (Präposition, Artikel-dat-fem-sg, Substantiv-nom-neut-sg, Substantiv-dat-fem-sg). Somit könnte das sächliche Wort `Amtsgericht` im erkannten Satz für das Sprachmodell grammatikalisch sinnvoll sein. Besonders schlecht wirken sich auch Komposita aus, die einen zusätzlichen Bindungskonsonanten (normalerweise `s` oder `n`) besitzen:

`zum Dreikönigstreffen` ⇒ `zum Drei Königs Treffen`
`zur Landesgartenschau` ⇒ `zur Landes Garten Schau`
`eine Straßenbahn` ⇒ `eine Straßen Bahn`

Neben dem Genus können dann auch Numerus und Kasus falsch interpretiert werden.

Als Konsequenz daraus, wurde für den zweiten Durchlauf das Sprachmodell sz1400K.arpa verwendet, welches aus dem grammatikalisch wesentlich korrekteren Text sz1400K (siehe auch Tabelle 5) erzeugt wurde. Als Referenz hierzu diente sz1400K.reg.arpa, also das Standardtrigrammmodell aus dem selben Trainingstext. In diesem Fall war das flexionsbasierte Modell mit 43,4 Prozent Fehlern leicht besser als das Standardmodell mit 43,8 Prozent. Alle Fehlerraten sind noch einmal in Tabelle 8 zusammengefasst.

[2] und natürlich auch inhaltlich sinnlos

	Fehlerrate, in Prozent
germ1M.1.arpa	45.4
germ1M.1.reg.arpa	44.2
sz1400K.arpa	43.4
sz1400K.reg.arpa	43.8

Tabelle 8: Fehler bei der Erkennung des Textes correct2K mit unterschiedlichen Sprachmodellen

Im Folgenden werden noch einige typische Fehler und Vorteile flexionsbasierter Sprachmodelle an Beispielen vorgestellt. Oft zu grammatikalischen Fehlern führen die unvollständigen Wörter (siehe 4.2.1). Da hierzu keinerlei Grammatikinformation vorliegt, kann das Sprachmodell im konkreten Fall nur 'raten'. Im Beispiel

korrekt:
`Nostalgie ist keine politische Tugend`
erkannt:
`Nostalgie es keine politischen Tugend`

ist das Wort `Tugend` unvollständig. Das Sprachmodell kann nicht wissen, dass es sich hierbei nicht um eine Pluralform handelt, die nach `keine politischen` grammatikalisch richtig wäre.

Ein anderes Problem ist, dass Verben oft nicht in unmittelbarer Nähe zum Subjekt des Satzes stehen, auf das sie sich beziehen. Die Verwendung von Trigrammmodellen erlaubt jedoch nur einen Rückblick auf die zwei letzten Wörter der Vergangenheit, wodurch die grammatikalische Form des Verbs für das Sprachmodell oft unklar bleibt.

korrekt:
`[...] bewaffnete Somalier acht Stunden lang bekämpft hatten`
erkannt:
`[...] bewaffnete Somalier acht Stunden lang gekämpft hat`

Für das Hilfsverb `haben` ganz am Ende des Satzes kann das Sprachmodell in diesem Beispiel alleine aus dem Bigrammkontext `lang bekämpft` nicht schließen, dass es sich auf mehrere Personen (`bewaffnete Somalier`) bezieht und verwendet daher fälschlicherweise die häufigere Singularform `hat`.

Allgemein sind jedoch inhaltlich falsch erkannte Sätze bei flexionsbasierten Modellen oft wenigstens grammatikalisch korrekt.

korrekt:
`[...] daß der Moderator den jüngsten der Bewerber fördert`

mit Standardmodell erkannt:
`[...] daß der Moderator den jüngsten der Bewerber für wert`
mit flexionsbasiertem Modell erkannt:
`[...] daß der Moderator den jüngsten der Bewerber verwehrt`

Das flexionsbasierte Modell verwendet mit `verwehrt` zumindest ein Verb in der korrekten grammatikalischen Beugung, auch wenn es hier keinen Sinn ergibt.

Interessant sind abschließend noch solche Fehler, die zeigen, dass die Unabhängigkeitsannahmen aus Abschnitt 2.2 nicht in allen Fällen streng gelten. Im Beispiel

korrekt:
`[...] kräftige Lackfarben wurden über das Land gegossen`
erkannt:
`[...] kräftige Leitfaden worden über das Land gegolten`

wird als letztes Wort `gegolten` verwendet, während das Standardmodell hier das korrekte `gegossen` wählt. Wahrscheinlich traten Sätze wie

```
Das Land gilt als sehr reich
Diese Länder gelten als Entwicklungsländer
```
etc.

häufig im Trainingstext auf. Nach Unabhängigkeitsannahme 1 (siehe 2.2.1) folgt also nach `das Land` oft eine Form von `gelten` unabhängig von der grammatikalischen Flektierung, hier also `gegolten`. Tatsächlich wurde wahrscheinlich das Trigramm (`das Land gegolten`) im Trainingstext nie gesehen.

5 Zusammenfassung

5.1 Bewertung des Ansatzes

Zusammenfassend lässt sich sagen, dass flekionsbasierte Sprachmodelle im Vergleich zu Standardmodellen zwar wesentlich höhere Perplexitäten aufweisen, in der realen Anwendung im Spracherkenner jedoch in etwa gleich gut abschneiden. Besonders wichtig für flexionsbasierte Modelle sind grammatikalisch einwandfreie Trainingstexte, insbesondere Komposita sollten nicht aufgetrennt werden. Zu den wichtigsten Fehlerursachen dieser Modelle zählen Wörter, für die keine grammatikalische Information vorliegt (unvollständige Wörter) und Verben. Zu den Stärken der Modelle gehören die Erzeugung grammatikalisch sinnvollerer Sätze und die Beherrschung unbekannter bzw. nicht gesehener Wortformen.

5.2 Zukünftige Erweiterungen und Verbesserungen

In Zukunft sollte daran gedacht werden, sich von dem Standardtrigrammmodell-Format für Sprachmodelle zu lösen. Dieses Format ist für flexionsbasierte Modelle denkbar ungeeignet, da es jedes Trigramm aus drei Wörtern (w_1 w_2 w_3) explizit aufgelistet verlangt. Ein Ansatz wäre, Grundform- und Flexionsoperatormodell auch im Erkenner getrennt zu lassen und erst bei Bedarf Wahrscheinlichkeiten nach der Grundformel (4) zu berechnen. Zur Zeit ist dies jedoch im Vergleich zur Vorberechnung noch zu zeitaufwändig.

Soll das Modell selbst verbessert werden, muss wohl zunächst über die Unabhängigkeitsannahmen nachgedacht werden. Da sie zentral zur Herleitung der Grundformel (4) beitragen, ist zu überlegen, ob diese tatsächlich in dieser Allgemeinheit gelten oder doch eingeschränkt werden müssen, was natürlich die beiden Teilmodelle vergrößern würde. Andererseits ist zu untersuchen, ob insbesondere für das Flexionsoperatorenmodell Trigramme ausreichen sind. Oft kann die Struktur des Satzes und damit die grammatikalischen Eigenschaften der einzelnen Wörter nicht aus einem Kontext der Größe 2 erkannt werden. Gerade im Fall der Verben treten deshalb viele Fehler auf. In Zukunft sollte man daher hierbei zu 4- oder 5-Grammen übergehen. Auch die Größe der Trainingstexte müsste für einen ernsthaften Einsatz im Spracherkenner wohl drastisch erhöht werden. Übliche Textgrößen von bis zu 100 Mio. Wörtern können jedoch im Moment aus zeit- und speicherplatztechnischen Gründen fast nicht durchgeführt werden. Dies liegt sicherlich zum einen an der Implementierung, die im Bezug auf Datenstrukturen und Algorithmen wohl noch verbessert werden könnte. Auf der anderen Seite belasten die durch die ungenaue Bestimmung entstehenden Ambiguitäten bei den Flektionsoperatoren die Performance erheblich, so dass zur Verarbeitung von sehr großen Texten wohl weitere Verbesserungen an der Tafel oder auch das Fallenlassen von weniger wahrscheinlichen Flektionsoperatoren nötig sind. Wünschenswert wären außerdem größere Grammatikwörterbücher, um Problemen aufgrund von unvollständigen Wortformen zu entgehen.

Auch über eine nachträgliche Verbesserung eines flexionsbasierten Modells kann in Zukunft nachgedacht werden: Zum einen wäre sicherlich eine Glättung mit einem der vielen Glättungsverfahren denkbar. Hierdurch wäre vor allem die geringe Größe des Trainingstexts abgemildert. Andererseits wäre auch eine Interpolation mit einem Standardtrigrammmodell denkbar, gerade im Hinblick auf die wesentlich geringen Perplexitäten, die diesen bieten.

Zu überlegen wäre auch ein ganz anderes Einsatzgebiet für ein solches Sprachmodell. Bei der Erkennung von Sprache erzeugt der Erkenner oft einen *Lattice*, also einen Graph möglicher Hypothesen für die sprachliche Eingabe, die jeweils einzeln durch ein Sprachmodell bewertet wurden. Ein solcher Latti-

ce könnte nun durch ein flexionsbasiertes Sprachmodell nachbewertet werden (*Rescoring*), um die grammatikalisch sinnvollen Hypothesen zu extrahieren und so den tatsächlich gesprochenen Satz zu finden.

Bei allen Unzulänglichkeiten flexionsbasierter Sprachmodelle, sind die jedoch wohl der aussichtsreichste Ansatz um in Zukunft Vokabulare in der Größenordnung von 1-2 Mio. Wörtern (typisch für die deutsche Sprache) implizit, also über Flexionsformen verwalten zu können. Standardtrigrammmodelle sind für solche Vokabulargrößen kaum sinnvoll.

Literatur

[1] CHEN, STANLEY F. und GOODMAN, JOSHUA, An Empirical Study of Smoothing Techniques for Language Modeling, Harvard University Cambridge, August 1998

[2] CLARKSON, PHILIP und ROSENFELD, RONALD, Statistical Language Modeling using the CMU-Cambridge Toolkit, 1994
`http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html`

[3] DRODOWSKI, GÜNTHER [Hrsg.], DUDEN 'Die Grammatik', Mannheim; Wien; Zürich: Bibliographisches Institut 1984

[4] JELINEK, FREDERICK; MERCER ROBERT L. und ROUKOS, SALIM, Principles of Lexical Language Modeling for Speech Recognition *in* FURUI, SADAOKI; SONDHI, M. MOHAN [Hrsg.] Advances in Speech Signal Processing, New York; Basel; Hong Kong, 1992

[5] LÜNGEN, HARALD; EHLEBRACHT, KARSTEN; GIBBON, DAFYDD und SIMÕES, ANA PAULA QUIRINO, Bielefelder Lexikon und Morphologie in VERBMOBIL Phase 2, November 1998

[6] RIES, KLAUS; SUHM, BERNHARD UND GEUTNER, PETRA, Language Modeling in JANUS, Oktober 1996

[7] ROSENFELD, RONALD, Statistical Language Modeling and N-Grams, Januar 1994