Learning, Forgetting and Remembering: Statistical Support for Rule-Based MT

Schriftliche Ausarbeitung meines Referats im Proseminar

„Einführung in die unifikationsbasierte maschinelle Übersetzung“

Munpyo Hong
WS 1999/2000

Garance Paris

14. April 2000

Sprach- und Translationswissenschaft / Maschinelle Übersetzung
Fachrichtung 5.6 Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen
Universität des Saarlandes

Contents

Einleitung

Von den verschiedenen Ansätzen in der maschinellen Übersetzung haben wir in diesem Seminar vor allem den sogenannten 'Rule-Based' Ansatz kennengelernt. 'Rule-Based' heißt, daß der Mensch versucht, linguistische Phänomene als Regeln innerhalb einer Grammatiktheorie zu formalisieren, und diese Regeln dann auf dem Rechner implementiert.

Diese Arbeit referiert über den Inhalt des Artikels

> ,,Learning, Forgetting and Remembering: Statistical Support for Rule-Based MT",
> in: TMI'99
> von Oliver Streiter, L. L. Iomdin, M. Hong & U. Hauck (1999).

Schwerpunkt des Artikels ist der Bericht über zwei Experimente, in denen ein Rule-Based System um vorher gesammelte statistische Information erweitert wurde.

Das Ziel einer solchen Integration mehrerer Ansätze ist ein besseres Ergebnis bei der Übersetzung, da die verschiedenen linguistischen Phänomene mit verschiedenen Strategien besser erfaßt werden können: Wortreihenfolge z. B. wird durch den Rule-Based Teil behandelt, Homographen aber werden mit Hilfe von Statistik unterschieden.

Zuletzt wird im Artikel das Sammeln von Information mit dem menschlichen Lernprozeß verglichen, und es wird ein Modell vorgeschlagen, wie man diesen Lernprozeß auf dem Rechner simulieren kann.

Erstes Experiment

Ziel der beiden Experimente ist das Anpassen des Systems auf den besonderen Bereich, 'subject domain', des zu übersetzenden Textes: Medizin, Sport, Wirtschaft, usw.

Im ersten Experiment wurden einsprachige Korpora benutzt. Das Sammeln von statistischen Information erfolgt automatisch, was schneller ist und als Vorteil hat, daß die Daten gleich in einem Format sind, die die Maschine weiterverwenden kann. Zum Sammeln dient ein Übersetzungsdienst auf dem Web, bei dem die Benutzer jeweils für ihren Text angeben müssen, zu welchem Bereich er gehört.

Zuerst wird eine morphologische Analyse der Texte durchgeführt, danach wird das Ergebnis davon statistisch analysiert, d. h. die Häufigkeiten der Lemmas werden gezählt. Die Ergebnisse der statistischen Analyse werden dann im Wörterbuch eingetragen.

Dieses Experiment wurde auf dem CAT2 System implementiert. Vor dem Zufügen von statistischer Information steht im Lexikon bei jedem Lemma der Quellsprache eine einzige Menge von möglichen Übersetzungen (s. Abbildung 1).

```
{l=allgemein,...,en={t=(common;general;generic;overall;public;universal)}}
{l=general,...,de={t=(allgemein;general;generell;gesamt;global)}}
```

Abbildung 1: ohne statistische Information

Nachdem das Lexikon um die statistische Information erweitert wurde, haben wir statt eine einzige Menge eine Disjunktion von Mengen, wobei jede Menge einer Domäne entspricht (s. Abbildung 2). Innerhalb jeder Domäne sind die Übersetzungsvorschläge als Liste geordnet, vom häufigsten zum weniger häufigen. Übersetzungen, die in der Domäne gar nicht gefunden wurden, werden nicht aufgezählt, und eine Domäne, in der keine der Übersetzungen gefunden wurde, wird nicht erwähnt (daher haben wir z. B. gar keine Übersetzung ins Englische für „allgemein" im Bereich Medizin).

Unterscheidung von Homographen

Solche statistische Information kann unter anderem dazu dienen, Homographen zu unterscheiden. Ursprünglich würden die Einträge für die ver-

```
{l=allgemein,...,
en=({t=(overall;universal;generic;general),dmn=sports}
   ;{t=(general;overall;generic;universal),dmn=society}
   ;{t=(general;universal;overall;generic),dmn=common}
   ;{t=(general;overall;generic;universal),dmn=comp_sc})}

{l=general,...,
de=({t=(gesamt;global;generell;general;allgemein),dmn=sports}
   ;({t=(gesamt;allgemein;generell;global;general),dmn=society}
   ;({t=(gesamt;allgemein;generell;global;general),dmn=common}
   ;({t=(allgemein;global;gesamt;generell;general),dmn=medicine}
   ;({t=(allgemein;gesamt;general;global;generell),dmn=economy}
   ;({t=(gesamt;allgemein;generell;general;global),dmn=comp_sc}))}
```

Abbildung 2: mit Domäne und statistischer Information

schiedenen Bedeutungen von „match" im Englischen wie in Abbildung 3 aussehen[1]. Bei solchen Einträgen gibt es aber keinerlei Möglichkeit zwischen den verschiedenen Einträge zu wählen.

```
{l=match,...; de={t=(uebereinstimmen;matchen)}}
{l=match,...; de={t=(streichholz)}}
{l=match,...; de={t=(spiel;partie)}}
```

Abbildung 3: ohne statistische Information

Mit statistischer Information erhält man aber die Einträge aus Abbildung 4. Bei der syntaktischen Analyse werden Übersetzungen, die in einer bestimmten Domäne nicht vorkommen, gar nicht betrachtet. Also werden in unserem Beispiel bei der Übersetzung von sportlichen Texten die Übersetzungen aus dem ersten Eintrag ausgeschlossen, da es hier keine Domäne „sports" gibt.

Eine solche Erweiterung des Lexikons hat aber auch ihre Grenzen, denn es könnte sein, daß Homographen, die zuerst unterschieden wurden, nachher nicht mehr unterschieden werden. Wenn im vorigen Beispiel „match" als verb-to-noun Derivation von „to match" in sportlichen Texten gefunden würde, also[1]

[1]Diese Einträge standen nicht wörtlich in dem Artikel, ich habe sie mir selbst zum besseren Verständnis rekonstruiert.

```
{l=match,...; de=({t=(uebereinstimmen;matchen),dmn=society}
                 ;{t=(uebereinstimmen;matchen),dmn=common}
                 ;{t=(uebereinstimmen;matchen),dmn=economy})}

{l=match,...; de=({t=(streichholz)})}

{l=match,...; de=({t=(spiel;partie),dmn=sports}
                 ;{t=(spiel;partie),dmn=society}
                 ;{t=(spiel;partie),dmn=common}
                 ;{t=(spiel;partie),dmn=economy}
                 ;{t=(spiel;partie),dmn=comp_sc})}
```

Abbildung 4: mit Domäne

```
{l=auto,...; en={t=(car;...;coach)}}
```

Abbildung 5: Übersetzungen für „Auto“, ohne statistische Information

```
{l=match,...; de=({t=(uebereinstimmen;matchen),dmn=sports})}
```

könnte man wiederum „übereinstimmen, matchen“ als Übersetzungen im sportlichen Bereich nicht mehr ausschließen.

Noise wegen Homographen in der Zielsprache

Dieses Problem läßt sich am besten an Hand eines Beispiels erklären. Das Wort „Auto“ übersetzt man ins Englische üblicherweise als „car“, aber wie Abbildung 5[1] zeigt, kann „coach“ auch vorkommen, obwohl selten. Außer im sportlichen Bereich, wo „coach“ doch häufiger ist (s. Abbildung 6[1]).

Wenn „coach“ im sportlichen Bereich häufiger vorkommt, liegt es daran, daß es auch Trainer bedeuten kann, und daß die Häufigkeit der beiden Homographen zusammen gezählt wird.

```
{l=auto,...; en=({t=(car;...),dmn=common},...,
                ;{t=(coach;car;...),dmn=sports})}
```

Abbildung 6: Mit Domäne und statistischer Information

```
{l=coach,...; de=({t=(bus;...),dmn=common})}
{l=coach,...; de=({t=(trainer;bus...),dmn=sports})}
```

Abbildung 7: „Zurückübersetzung“ von „coach”

Um eine falsche Übersetzung zu vermeiden, schaut man sich die andere Übersetzungsrichtung an. Für die Übersetzung von „coach“ werden zwei verschiedene Einträge gefunden. Es handelt sich um Homographen, die normalerweise nur in verschiedenen Bereichen vorkommen, mit völlig verschiedenen Bedeutungen und Übersetzungen. Dieses merkt man daran, daß die Übersetzung von „coach“ im ersten Eintrag, also „Bus“, im Bereich Sport gar nicht vorkommt (in Abbildung 7[1] durchgestrichen).

Daraus schließt man, daß im deutschen Lexikon die Häufigkeiten der beiden Homographen gemischt worden sind, und daß die erste Übersetzung nur irrtümlich diese hohe Häufigkeit bekommen hat. Daher wird sie gar nicht verwendet, sondern es wird die zweite Übersetzung gewählt.

Ergebnisse

Auswirkungen auf die Parsingdauer

Um die Auswirkungen auf die Parsingdauer einzuschätzen, wurden 50 Sätze einmal mit und einmal ohne Häufigkeitsinformation geparst. Bei 17 Sätzen war das Ergebnis mit Einberechnung der Häufigkeiten positiv, sonst wurde kein Unterschied beobachtet. Insgesamt war die Dauer um 13 % gesunken.

Qualität der Übersetzung

Das erweiterte System wurde getestet, indem Sätzen übersetzt wurden, und die Übersetzung von menschlichen Übersetzern als besser oder schlechter eingeschätzt wurde. In 7 von 40 Fällen war das Ergebnis besser als ohne statistische Information, in zwei Fällen wurde sie aber schlechter. Dies liegt daran, daß es in den Daten keine Beziehung zwischen Quell- und Zielsprache gibt: es wird nicht gezählt, wie oft ein Wort mit einem anderen übersetzt wurde, sondern wie oft eine mögliche Übersetzung in der Zielsprache vorkommt.

Zweites Experiment

Um bessere Ergebnisse zu bekommen, hat man also in dem zweiten Experiment versucht, diese Beziehung mit zweisprachigen Korpora (Korpora, in denen jeweils ein Text mit seiner Übersetzung vorhanden ist) wiederzugeben.

Aus der morphologischen Analyse der beide Texten bekommt man einen String von Lemmas. Dann wird von einem Programm berechnet, welche Teile im Ausgangstext welchen Teilen im Zieltext entsprechen. So kann man für jedes Wort im Ausgangstext einfach erkennen, was es für Übersetzungsmöglichkeiten im Lexikon gibt, und welche von diesen Möglichkeiten tatsächlich in der Übersetzung in einem bestimmten Bereich erschienen sind.

Diese Übersetzungen mit ihren Häufigkeiten, jeweils auf eine Domäne bezogen, werden dann im Lexikon eingetragen. Da die Häufigkeiten in diesem Fall die Häufigkeiten der Übersetzungspaare sind und nicht der übersetzten Wörter selbst, würde man erwarten, daß dieser Ansatz bessere Ergebnisse liefere, aber evaluiert wurde er noch nicht.

Conclusion: Lernen, Vergessen, Erinnern

Das Sammeln von Daten kann man mit einem Lernprozeß vergleichen. Dieser Lernprozeß muß aber von einem Prozeß des Vergessens unterstützt werden, der dafür sorgt, daß die Daten aktuell bleiben, und daß Fehler vom Speicher entfernt werden. Dieser Prozeß des Vergessens kann einfach implementiert werden, indem alte Daten, die niedrige Häufigkeiten haben, angegriffen werden, und ihre Häufigkeiten reduziert werden.

Dazu braucht man aber auch noch einen Erinnerungsprozeß, der erlaubt, einmal vergessene Daten zu einem späteren Zeitpunkt schneller zu lernen. Dies wird mit einem passiven Speicher implementiert.

Dieser Lernprozeß wird in Abbildung 8 graphisch dargestellt. Wenn ein Element im Korpus gefunden wird, wird es in den passiven Speicher eingetragen. Die Häufigkeit alter Elemente, deren Häufigkeit einen bestimmten Schwellenwert nur wenig überschreitet, wird um eins reduziert. Nur Wörter, deren Häufigkeit den Schwellenwert immer noch überschreitet, werden in den aktiven Speicher kopiert und stehen dann für die Übersetzung zur Verfügung. Bei den restlichen Wörtern wird die Häufigkeit um eins reduziert. Also bleiben Wörter, die nicht häufig vorkommen, im passiven Speicher: Sie sind ,,vergessen“ worden, aber können später leichter nochmals gelernt werden, man kann sich sozusagen daran ,,erinnern“.

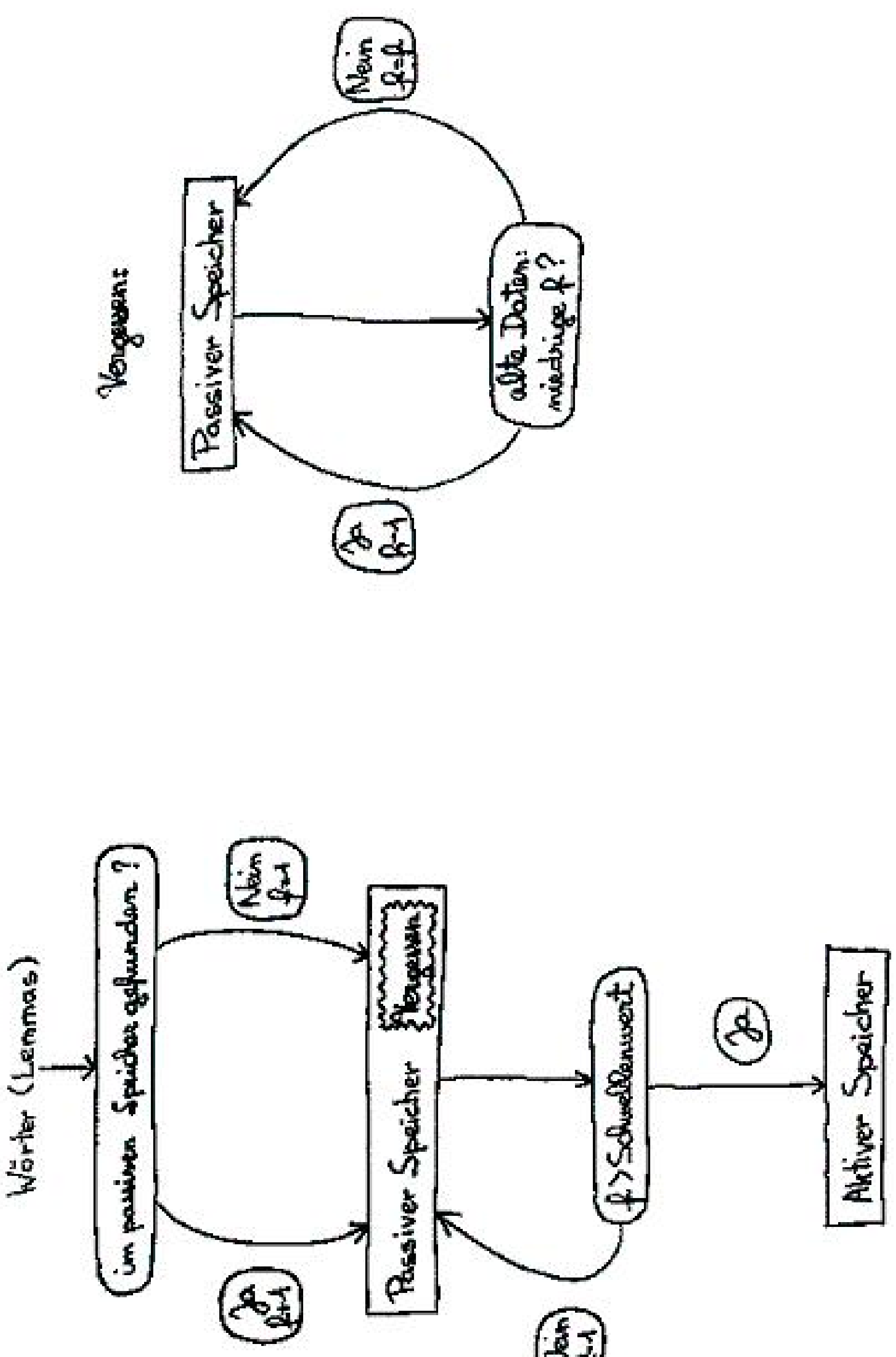

Abbildung 8: Lernen, Vergessen, Errinern