Spektrum.de

01.12.1996

Linguistische Analyse in einem multilingualen Sprachsynthese-System

Bernd Möbius und Richard Sproat

Das maschinelle Umwandeln von geschriebenem in gesprochenen Text (fachlich kurz text to speech, TTS) besteht prinzipiell aus zwei Schritten: der linguistischen Analyse und der akustischen Sprachsynthese. Hier geht es im wesentlichen um den ersten Schritt: Aus dem schriftlichen Eingabetext ist eine linguistische Repräsentation herzuleiten und für die Synthese bereitzustellen. Dieses Zwischenstadium ist typischerweise eine Kette von Lautsymbolen zusammen mit Informationen über Phrasierung, Sprachmelodie und Betonung.

Betrachten wir beispielsweise den deutschen Satz "Bei der letzten Wahl gewann John Major ca. 42% der Wählerstimmen." Welches Wissen muß ein Sprecher des Deutschen mitbringen, um ihn korrekt vorzulesen?

Zunächst muß er die Aussprache regulärer deutscher Wörter aus ihrer schriftlichen Form ableiten können. Dies setzt unter anderem die Kenntnis der inneren Struktur von Wörtern voraus. So muß er "Wählerstimmen" in die Komponenten "Wähler" und "Stimmen" zerlegen, um die Buchstabenfolge "st" korrekt als "scht" auszusprechen, im Unterschied etwa zu dem Wort "Erstimpfung". Des weiteren sollte er "John Major" als

ausländischen Namen erkennen und idealerweise auch englisch aussprechen. Die Abkürzungen "ca." und "%" sowie die Zahl "42" schließlich sind vor jeder weiteren Aktion in reguläre Wortformen umzuwandeln. Eine Besonderheit ist dabei, daß der Punkt im einen Falle eine Abkürzung, im anderen das Satzende markiert. Schließlich muß der Sprecher noch Wörter und Silben richtig betonen und dem Satz eine geeignete Sprachmelodie geben (siehe dazu den Beitrag von Klaus Kohler).

In den Bell Laboratories in Murray Hill (New Jersey) haben wir in den vergangenen Jahren ein multilinguales TTS-System entwickelt. Derzeit gibt es Versionen für Englisch, Französisch, Spanisch, Italienisch, Deutsch, Russisch, Rumänisch, Chinesisch und Japanisch. Die zugrundeliegende Software sowohl für die linguistische Analyse wie für die Sprachausgabe ist für alle Sprachen identisch. Das System benötigt zwar Information über die verschiedenen Schriftsysteme (zur Zeit verfügbar für lateinische, kyrillische, chinesische und japanische Schrift) sowie über das akustische Inventar und spezielle Regeln für die linguistische Analyse jeder Sprache; diese sind jedoch in externen Tabellen und Dateien abgelegt, auf die das Programm erst zur Laufzeit zugreift.

Das System kann deshalb ohne weiteres Stimme und Sprache wechseln. Das ist vor allem interessant beim Vorlesen von Dialogen oder elektronischer Post. Der Aufbau ist vergleichbar dem eines üblichen Textverarbeitungsprogramms mit einem sprachunabhängigen Kern und sprachspezifischen Zeichensätzen, Trennungsregeln, Konventionen für die Schreibweise des Datums und so weiter.

Kehren wir noch einmal zu unserem Beispielsatz zurück. Ein

noch nicht erwähntes Problem der Textanalyse ist die Zerlegung des Eingabetextes in Wörter. Dies ist selbst für das Deutsche, das Wörter in der Regel durch Leerzeichen abtrennt, keine triviale Aufgabe. So umfaßt die Zeichenfolge "42%" die zwei separaten Wörter "zweiundvierzig" und "Prozent". Wesentlich schwieriger verhält es sich in Sprachen wie dem Chinesischen oder Japanischen, in denen man Wörter überhaupt nicht an Kennzeichen im Schriftbild erkennt. Dennoch existieren in diesen Sprachen Wörter als lexikalische Einheiten, so daß die linguistische Analyse auch hier eine Wortsegmentierung vornehmen muß (Bild 1). Das wörtliche Ausschreiben eines Symbols wie "%" ist in einigen Sprachen ebenfalls komplexer als im Deutschen, wo es ausnahmslos durch "Prozent" zu ersetzen ist. Dagegen kann es im Russischen je nach Satzzusammenhang ein Substantiv oder ein Adjektiv sein, das auch noch korrekt zu deklinieren ist (Bild 2).

Eine Besonderheit des Deutschen wiederum sind Komposita (zusammengesetzte Wörter) wie "Wählerstimmen". Da die Bildung neuer Zusammensetzungen üblich ist, treten in nahezu jedem Text welche auf, die in keinem noch so umfangreichen Lexikon verzeichnet sind. Die linguistische Analyse muß darum Komposita in ihre Bestandteile zerlegen können. Der vielzitierte "Donaudampfschiffahrtsgesellschaftskapitän" ist dabei für die Praxis weniger typisch als ein Wort wie "Un-anständig-keit-s-unterstellung". Bereits dieses muß in die durch Bindestriche angedeuteten Komponenten zerlegt werden, damit man sie alle im Lexikon wiederfindet.

Die Vielfalt dieser Probleme scheint zunächst gegen eine generelle Lösung zu sprechen. Unter einer abstrakteren allgemeinen Betrachtungsweise wird jedoch ihre gemeinsame

Struktur erkennbar: Jedes Teilproblem ist beschreibbar als die Aufgabe, eine Kette von Symbolen in eine andere zu verwandeln, zum Beispiel (erster Schritt) den geschriebenen Text "Wählerstimmen" in eine linguistische Repräsentation, die nun auch Informationen über die Struktur des Wortes enthält: "wähler{substantiv}+stimme{substantiv} +n{plural}". Auf vergleichbare Weise wird aus einer Folge von Schriftzeichen in einem chinesischen Satz eine Darstellung, die unter anderem Informationen über Wortgrenzen enthält.

Der nächste Schritt, die Bestimmung der Aussprache von Wörtern, baut auf solchen Informationen auf. Mit Hilfe der Ausspracheregeln für eine bestimmte Sprache wird die linguistische Repräsentation in eine Folge von Lautsymbolen konvertiert. Nur weil bereits im ersten Schritt die wortinterne Grenze vor "st" in "Wählerstimmen" richtig gesetzt wurde, kann der zweite die korrekte Aussprache des Wortes bestimmen.

Ein flexibles und zugleich mathematisch elegantes Modell für die skizzierte Konvertierung von Symbolketten ist der finite state transducer (FST): eine abstrakte Maschine, die eine endliche Anzahl von Zuständen annehmen kann. Sie nimmt ein Eingabesymbol entgegen, geht daraufhin in einen anderen (oder denselben) Zustand über und gibt ein Symbol aus, und zwar nach Maßgabe einer Tabelle, die für jeden Zustand und für jedes Eingabesymbol den anzunehmenden Zustand und das Ausgabesymbol vorschreibt (Kasten auf Seite 104). Für eine komplexe Aufgabe wie die linguistische Analyse in einem Sprachsynthese-System muß ein FST typischerweise einige hunderttausend mögliche Zustände haben.

Ronald Kaplan und Martin Kay vom Forschungszentrum der Firma Xerox in Palo Alto (Kalifornien) haben in den siebziger

Jahren erstmals FSTs auf linguistische Problemstellungen angewandt. Sie zeigten auch, daß man von Experten vorab erstellte linguistische Beschreibungen, etwa Ausspracheregeln, automatisch in FST-Maschinen konvertieren lassen kann. Tradition hat diese Forschungsrichtung insbesondere an den Universitäten Helsinki und Paris VII.

In unseren Arbeiten verwenden wir eine spezielle Variante von Transducern, die gewichteten FSTs, deren Grundlagen unsere Institutskollegen Michael Riley, Fernando Pereira und Mehryar Mohri erforscht haben. Für eine bestimmte Kombination aus Zustand und Eingabesymbol stehen einem gewichteten FST mehrere Alternativen offen. Jede trägt eine Art Preisschild (man spricht von Gewichten); ein Übergang in einen neuen Zustand samt zugehörigem Ausgabesymbol wird als um so teurer deklariert, je weniger wahrscheinlich er für sich genommen ist. Die Maschine verfolgt dann mehrere Alternativen parallel weiter und entscheidet sich schließlich für die Folge von Übergängen, die den günstigsten Gesamtpreis und damit die größte Plausibilität für sich hat.

Die linguistische Analysekomponente in unserem TTS-System ist vollständig nach diesen Prinzipien konstruiert. Die dadurch erzielte einheitliche Software-Architektur ermöglicht eine vergleichsweise einfache Erweiterung auf neue Sprachen, und ihre modulare Struktur erleichert die Integration verbesserter Komponenten in bereits existierende Systeme.

Aus: Spektrum der Wissenschaft 12 / 1996, Seite 103

Bernd Möbius und Richard Sproat

Möbius entwickelte die deutsche Version des beschriebenen TTS-Systems|; seine besonderen Interessen sind Phonetik und Prosodie. Er arbeitet zusammen mit Dr. Sproat in der Abteilung für Sprachsynthese-´Forschung der Bell Laboratories in Murray Hill (New

Jersey). Die Bell Laboratories sind die Forschungseinrichtungen von Lucent Technologies, der Kommunikationstechnologie-Firma, die aus dem Telekommunikations-Konzern AT&T ausgegliedert wurde. Weitere Informationen sowie eine interaktive Demonstration des TTS-Systems der Bell Labs finden sich im WWW unter http|://www.bell-labs.com/project /tts/. Sproat entwarf das Design für multilinguale Textanalyse und entwickelte ein Toolkit zur lexikalischen Analyse|; seine Spezialgebiete sind Computerlinguistik und Morphologie. Er arbeitet zuammen mit Dr. Möbius in der Abteilung für Sprachsynthese-Forschung der Bell Laboratories in Murray Hill (New Jersey). Die Bell Laboratories sind die Forschungseinrichtungen von Lucent Technologies, der Kommunikationstechnologie-Firma, die aus dem Telekommunikations-Konzern AT&T ausgegliedert wurde. Weitere Informationen sowie eine interaktive Demonstration des TTS-Systems der Bell Labs finden sich im WWW unter http|://www.bell-labs.com/project/tts/.