Testmethoden für die sichere, sinnentsprechende Silbentrennung und andere Anwendungen einer Wortanalyse

Abstract

Schlagwörter: Testmethoden, Wortanalyse, Silbentrennung, sinnentsprechende Volltextsuche, neue deutsche Rechtschreibung

Sichere sinnentsprechende Silbentrennung
In diesem Aufsatz wird ein umfassendes Testsystem für die sichere, sinnentsprechende Silbentrennung [1] vorgestellt. Diese Anwendung basiert auf der Wortanalyse: dabei werden zusammengesetzte Wörter in ihre Einzelwörter und diese wiederum gemäß der deutschen Wortbildungsgrammatik in ihre atomaren Wortbestandteile (=Atome) zerlegt. Die Atome können nach ihrer Verwendung bei der Wortbildung grob in die Klassen Vorsilbe, Stamm und Endung eingeteilt werden. In der Praxis hat es sich jedoch bewährt, eine detailliertere Klassifizierung nach Wortarten zu verwenden und so vor allem die Menge der erlaubten Endungen für bestimmte Stämme einzuschränken [3]. Das Wortanalysesystem besteht aus zwei wesentlichen Bestandteilen: der Atomtabelle, die alle atomaren Wortbestandteile (=Atome) enthält, die in bestimmte Klassen eingeteilt sind, und dem Algorithmus zur Zerspaltung der Wörter in ihre Atome gemäß den angegebenen Grammatikregeln. Nach der Zerlegung hat man bereits die Trennstellen zwischen Einzelwörtern und hinter Vorsilben gefunden. Mit Hilfe einer weiteren Methode werden die übrigen Trennstellen in dem noch ungetrennten, aus Stamm und Endung(en) bestehenden Teil auf Basis der Vokal-Konsonanten-Folgen gesucht.

Testsystem
Durch umfangreiche Tests der Silbentrennung soll sichergestellt werden, dass korrekt geschriebene Wörter garantiert keine falsche Trennstelle enthalten und grammatikalisch oder orthographisch fehlerhafte Wörter nach Möglichkeit erkannt werden.

Im Idealfall würden alle existierenden Wörter getestet und kontrolliert. Dies ist aufgrund der Möglichkeit zur ständigen Schaffung neuer Wörter und aufgrund des hohen Aufwandes nicht möglich. Zielführender sind dagegen selektive Tests, die dem Benutzer nur jene Wörter zur Kontrolle liefern, die bestimmte Kriterien erfüllen.

Ein Testsystem wurde implementiert, das die Wortanalyse anhand der konkreten Anwendung der Silbentrennung testet. Das System integriert verschiedene Versionen der Wortanalyse: zwei Verfahren verwenden eine sehr detaillierte Klassifizierung der Atome unter Berücksichtigung der deutschen Wortbildungsgrammatik, eines davon in Verbindung mit den seit 1998 gültigen neuen Rechtschreibregeln, das andere mit den alten Rechtschreibregeln. Zum Vergleich wurde auch ein primitives Verfahren mit einfachster Wortgrammatik eingebaut, das Wörter nach den alten Rechtschreibregeln erkennen und trennen kann, jedoch für viele Wörter mehrere, oft auch nicht sinnvolle Zerlegungen zulässt.

Das vorgestellte Testsystem ermöglicht auf einfachste Weise das Testen der Silbentrennung auf der Basis der Wortanalyse. Es können sowohl einzelne Wörter als auch komplette Textdateien in unterschiedlichen Textformaten getestet werden. Die Testergebnisse werden in übersichtlicher Form sowohl am Bildschirm ausgegeben als auch in eine Ausgabedatei geschrieben.

Als Ergebnis der Silbentrennung liefert jedes Verfahren für das Eingabewort alle gefundenen Trennvarianten des Wortes. Trennstellen zwischen Einzelwörtern werden durch Haupttrennstellen (=), solche innerhalb von Einzelwörtern durch Nebentrennstellen (-) gekennzeichnet.

Unterschiedliche Trennungen für ein Wort resultieren dabei oft aus der Tatsache, dass es für dieses Wort mehrere, grammatikalisch korrekte, aber nicht immer auch sinnvolle Zerlegungen gibt (z.B. *Per-so-nal=man-gel*, *Per-son=alm=an-gel*).

Das Testsystem erlaubt sowohl den Vergleich der Resultate der drei Verfahren als auch den intensiven Test jedes einzelnen Verfahrens. Im Rahmen des selektiven Tests wurden für die beiden auf der deutschen Wortbildungsgrammatik basierenden Verfahren besondere Kriterien erarbeitet. Diese Kriterien wurden nach solchen Gesichtspunkten ausgewählt, dass die wenigen potentiellen Problemfälle gezielt aufgedeckt werden. Beispielsweise können Wörter herausgefiltert werden, die Atome mit unterschiedlicher Funktionalität enthalten und deshalb problematisch sind (z.B. „ende“ kann als Stamm oder als Endung verwendet werden: das *Spiel=en-de* bzw. *spie-len-de* Kinder). Umfangreiche Tests mit Hilfe dieser Testumgebung trugen entscheidend zur Verbesserung des Wortanalysesystems bei, indem zum Beispiel Fehler bei der Klassifizierung der Stämme oder fehlende Stämme, z.B. selten benötigte Fremdwörter, entdeckt und diese Mängel in der Folge behoben werden konnten.

Weitere Anwendungen
In das Testsystem können auf einfache Weise andere Anwendungen integriert werden, die auf Wortanalyse basieren und auf die beschriebene Art getestet werden sollen. Ein Beispiel dafür ist die sinnentsprechende Volltextsuche, die durch Anwendung des Zerlegungsalgorithmus für jedes Eingabewort eine Menge der sinngebenden Bestandteile liefert [2]. Sinngebende Bestandteile sind in Wortzusammensetzungen die Einzelwörter, in Einzelwörtern die Vorsilben in Verbindung mit den Stämmen (z.B.: *Wortzerlegungsverfahren* $\rightarrow$ {*wort, zerlegung, verfahren; zerleg, verfahr*}). Werden mehrere Zerlegungen für ein Wort gefunden, so lässt sich der Sinn des Wortes oft nicht mehr eindeutig feststellen. Das Resultat enthält dann mehrere Mengen von sinngebenden Bestandteilen (z.B. *Baumast* $\rightarrow$ {*baum, ast*}, {*bau, mast*}). Im Bereich der sinnentsprechenden Suche soll durch Testen gewährleistet werden, dass die Menge der sinngebenden Wortbestandteile für alle Wörter korrekt identifiziert wird. Eine Funktion, die für alle drei Verfahrensversionen die sinngebenden Bestandteile als Ergebnis ausgibt, wurde bereits in das Testsystem integriert. Es ist vorgesehen, solche Funktionen auch für die Rechtschreibprüfung und die Groß- und Kleinschreibung zur Verfügung zu stellen.

Literatur

[1] Barth, W., Nirschl H.: Sichere sinnentsprechende Silbentrennung für die deutsche Sprache. Angewandte Informatik 4, S. 152-159, 1985.

[2] Barth, W.: Volltextsuche mit sinnentsprechender Wortzerlegung. Wirtschaftsinformatik, 32. Jahrgang, Heft 5, S. 467-471, 1990.

[3] Steiner, H.: Automatische Silbentrennung durch Wortbildungsanalyse. Dissertation, Institut für Computergraphik, Technische Universität Wien, 1995.