34 Sprachen durch Zählen von Wörtern bändigen

Dank einer mathematischen Gesetzmäßigkeit in der Sprache kann Google im Bruchteil einer Sekunde eine Suchanfrage beantworten oder einen Text automatisch übersetzen.

Welches ist das am häufigsten verwendete Wort im Niederländischen? Das hängt davon ab, um welche Art von Sprache es sich handelt. Wenn es um das geschriebene Niederländisch in Zeitungen und Zeitschriften geht, ist „das“ das am häufigsten verwendete Wort. Nimmt man jedoch das gesprochene Niederländisch, erscheint oben das Wort „Ja“. Und auf Twitter ist „Ich“ der Spitzenreiter.

Doch das geschriebene Niederländisch, das gesprochene Niederländisch und das Twitter-Holländisch haben eines gemeinsam: Das häufigste Wort innerhalb einer solchen Domäne ist doppelt so häufig wie Nummer zwei auf der Rangliste, dreimal so häufig wie Nummer drei und so weiter. Wenn wir die Frequenz des häufigsten Wortes auf 1 setzen, dann bilden die Worthäufigkeiten die Reihe 1, 1/2, 1/3, 1/4, ...

Dieses in einer mathematischen Formel ausgedrückte Muster wird *Zipfsches Gesetz* genannt, nach dem amerikanischen Linguisten George Zipf, der das Gesetz 1935 entdeckte. „Dieses Gesetz scheint für alle Sprachen und für alle Textsammlungen innerhalb einer Sprache zu gelten, egal ob man in einem chinesischen Rechtsbuch, einer norwegischen Bibel oder in englischsprachigen E-Mails eines großen Unternehmens nachschlägt“, sagt Antal van den Bosch, Professor an der Radboud Universität Nimwegen und Spezialist für Computerlinguistik. „Das Zipfsche Gesetz ist ein empirisches Gesetz, aber es ist ziemlich genau. Nur am Anfang, bei den Top 10 der Wörter, und am Ende, bei den seltenen Wörtern, weicht die Praxis ein wenig von der mathematischen Formel ab.“

Effiziente Suche

Gerade weil das Gesetz von Zipf allgemeingültig ist, kann die Suchmaschine von Google so schnell antworten. Van den Bosch: „Der Trick von Google besteht darin, dass sie einen Wortindex des Webs erstellt haben und diesen ständig aktualisieren. Der Wortindex zeigt an, welches Wort in welchem Dokument enthalten ist. Mit dem Zipf-Gesetz können Sie nun zeigen, dass dieser Wortindex kompakt ist. Und das wiederum bedeutet, dass Sie es kompakt auf Festplatten speichern und problemlos an Datenzentren auf der ganzen Welt verteilen können."

Warum genau ist dieser Wortindex kompakt? Google hat Zugang zu Dutzenden Milliarden von Webseiten, aber die Anzahl der Wörter pro Sprache liegt „nur" im Millionenbereich, von denen es in einem offiziellen Wörterbuch normalerweise nur einige Hunderttausend gibt. Das Zipf-Gesetz lehrt uns nun, dass die Hälfte der Wörter in einer großen Textsammlung nur einmal vorkommt. Dank Zipf wissen wir auch, dass die Top 300 fast alle Funktionswörter (Artikel, Pronomen, Präpositionen ...) und die am häufigsten verwendeten Inhaltswörter (Substantive, Verben, Adverbien, Adjektive) enthalten. Diese beiden Eigenschaften machen den Wortindex kompakt.

Van den Bosch: „Wenn wir einen Suchbegriff eingeben, muss Google nicht in Zehnmilliarden von Dokumenten suchen, sondern in dem viel handlicheren Wortindex. Und wenn jemand vier Schlüsselwörter eingibt, nimmt die Suchmaschine die Überschneidung von vier Sammlungen. Jeder Satz sagt aus, auf welcher Webseite das Wort erscheint. Es ist eine einfache Berechnung, also geht es schnell."

Automatische Übersetzung

Maschinelle Übersetzungsmaschinen nutzen eine Art abgeleitete Eigenschaft des Zipf-Gesetzes aus, nämlich die Eigenschaft, Wortkombinationen zu verhindern. Google Translate verwendet eine große Datenbank mit bereits vorhandenen Übersetzungen, z.B. offiziell übersetzte Texte des Europäischen Parlaments oder übersetzte Untertitel von Filmen. Um einen neuen Text z.B. vom Niederländischen ins Englische zu übersetzen, sucht die Übersetzungsmaschine nach möglichst langen Wortkombinationen, die in bestehenden Übersetzungen so oft wie möglich auf die gleiche Weise übersetzt wurden.

Die Übersetzungsmaschine sieht zum Beispiel, dass das Shakespeare-Zitat „'Juliet is the sun" immer mit „Julia ist die Sonne" übersetzt wird. In diesem Fall muss es sich um die korrekte Übersetzung handeln. Wie oft Wortkombinationen in einer bestimmten Reihenfolge vorkommen, wird ebenfalls in Zipf-ähnlicher Weise aufgeteilt: nur eine begrenzte Anzahl von Kombinationen ist sehr häufig. Und so wie der Wortindex kompakt ist, so ist auch der Index der Wortkombinationen kompakt. Deshalb erledigt eine Übersetzungsmaschine ihre Arbeit so schnell.

Diese statistische Übersetzungsmethode eignet sich gut für Texte, die bestehenden Texten sehr ähnlich sind. Aber je einzigartiger und kreativer der Text ist, desto schwieriger ist es für die Übersetzungsmaschine. „Poesie ist notorisch schwierig", sagt van den Bosch. „Der heilige Gral auf meinem Gebiet lautet daher: Wie können wir sicherstellen,

dass Maschinen Sprache wirklich verstehen? Denn Google Translate macht das immer noch nicht, egal wie nützlich es oft ist.“