Text Summarization - Eine Übersicht

Philip Kurmann, philip@kman.ch

18 Juni 2017

Einführung

Durch das Internet, welches in den späten 1990er Jahren seinen Durchbruch schaffte, ist die Anzahl veröffentlichter Texte und Dokumente explodiert. Man ist heute einer Fülle von Informationen ausgesetzt, welche kein Mensch mehr bewältigen kann und es wird immer schwieriger, effizient und effektiv die richtigen Informationen zu finden.

Ein Ansatz, um mit dieser Informationsflut besser umgehen zu können, ist die Erstellung, von automatischen Textzusammenfassungen. Dabei geht es darum, die Komplexität und die Länge eines Textes zu reduzieren, bei gleichzeitiger Erhaltung der essentiellen Informationen des original Dokuments. Erste Forschungen zum Thema Automatische Textzusammenfassung begannen in den 1960er Jahren. Durch das enorme Wachstum an Informationen und Dokumenten wurde das Interesse an solche Systemen jedoch immer grösser. In jüngster Zeit wird intensiv an diesem Thema geforscht.

Automatische Textzusammenfassung kann für folgende Aufgaben hilfreich sein:

- Übersichtliche Darstellung von News auf kleinen Bildschirmen (Bsp.: Handys)
- Vorlesen eines Textes mittels Sprachsynthese (Bsp. Siri, Cortana oder Alexa). Der komplette Text ist u. U. zu langwierig.
- Eine kurze und prägnante Zusammenfassung des Suchresultates von Suchmaschinen (Bsp. Google)
- Übersetzungen - Den Text vor der Übersetzung kürzen.

Was ist Text Summarization?

Textzusammenfassung kann folgendermassen definiert werden: *Ein Text, der aus einem oder mehreren Texten produziert wurde, welcher einen signifikanten Anteil an Informationen der Originaltexte beinhaltet und welcher nicht mehr als die halbe Länge der Originaltexte ist.* (Hovy 2005)

Wenn nun eine Textzusammenfassung automatisch durch einen Computer erstellt wird, nennt man das Verfahren *automatische Textzusammenfassung.* Dabei kann die Zusammenfassung entweder aus einem Dokument erstellt worden sein, oder aus mehreren. Bei mehreren heisst das Verfahren *Multi Document Summarization (MDS).* Gleichzeitig können Dokumente von einer Sprache (*monolingual*) oder von mehreren unterschiedlichen Sprachen gleichzeitig zusammen gefasst werden (*translingual oder multilingual*).

Man unterscheidet zwischen einem *extract*, bei dem (Teil-)Sätze und Textstellen aus den Originaldokumenten übernommen werden und einem *abstract*, bei dem die Inhalte mit neuen Sätzen zusammengefasst werden.

Extraktion vs. Abstraktion

Erstellt ein Mensch eine Zusammenfassung eines Textes, verwendet er dafür grösstenteils eigen Wörter und Sätze. So funktionieren professionelle Unternehmen, welche Textzusammenfassungen erstellen (Bsp.: Cliffs Notes).

Automatische Textzusammenfassung basiert auf statistischen, linguistischen und heuristischen Methoden. Das System berechnet, wie oft gewisse Schlüsselwörter in den Originaltexten vorkommen. Dabei werden die Frequenz der Schlüsselwörter in den Texten, in welchen Sätzen diese vorkommen und wo diese Sätze sich in den Originaltexten befinden berücksichtigt, um die Schlüsselstellen des Textes zu ermitteln. Zusätzlich können auch Textformatierungen berücksichtigt werden.

Textextraktion ist sehr viel einfacher als Textabstraktion. Für die Extraktion müssen die Schlüsselstellen ermittelt werden, was mittels empirischen Methoden oder Maschine Learning durchgeführt werden kann. Für die Abstraktion braucht es jedoch zusätzlich ein System, das den Text versteht und Wissen besitzt, wie daraus ein neuer Text geschrieben werden kann. Die meisten automatischen Textsummarization Tools generieren daher nur Extracts und keine Abstracts.

Methoden

Es gibt verschiedene Arten von Textzusammenfassungen. Diese werden bestimmt, durch die Anzahl Dokumente, welche zusammengefasst werden und die Anzahl Sprachen.

Bei der *Single-Document Summarization* wird einer Zusammenfassung eines einzelnen Dokumentes erstellt. Dies ist die einfachste Methode, da alle Informationen innerhalb des Dokumentes als aktuell bewertet werden könne.

Bei der *Multi-Document Summarization* wird eine Zusammenfassung über mehrere Dokumente erstellt, welche in der selben Sprache verfasst sind. Hier werden Dokumente von verschiedenen Quellen für ein Thema zusammengefasst. Eine Schwierigkeit besteht bei dieser Methode darin, dass die Dokumente in den richtigen zeitlichen Kontext gebracht werden müssen. Denn nicht alle Informationen sind gleich aktuell und typischerweise ersetzen neuere Informationen die Älteren. Um die richtigen Aussagen zu machen, muss daher die Zeit berücksichtigt werden, was nicht immer ganz einfach ist. Eine weitere Schwierigkeit bei *multi-Document Summarizations* besteht darin, thematische Überlappungen von den einzelnen Dokumenten zu identifizieren, um Wiederholungen auszuschliessen. Sehr schwierig wird es jedoch, wenn sich die Dokumente inhaltlich widersprechen. Ein Ansatz kann sein, dass man mit Hilfe des zeitlichen Kontextes aktuelleren Inhalt höher gewichtet.

Eine weitere Methode ist die *Multi-Lingual Summarization*. Grundsätzlich kann man diese Methode als eine Erweiterung der *Multi-Document Summarization* betrachten. Im Gegensatz dazu, sind die Dokumente jedoch in unterschiedlichen Sprachen verfasst.

Die 3 Phasen der automatischen Textzusammenfassung

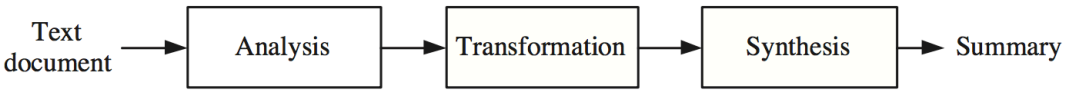

Figure 1: Prinzip Textzusammenfassung (Choon-Ching und Selamat 2014, p. 1)

Abbildung 1 stellt das Prinzip der Textzusammenfassung dar. Der Prozess kann in drei Phasen unterteilt werden: Analyse, Transformation und Synthese.

Phase 1: Analyse oder Topic Identification

Die Analyse Phase analysiert den Text und extrahiert wichtige Teile. Sie identifiziert die Grundlegenden Wörter, Sätze und Abschnitte eines Textes. Typischerweise werden in dieser Phase mehrere komplementäre Methoden verwendet. Heutige Systeme beschränken sich mehrheitlich auf diese Phase (siehe Hovy 2005).

Um die Analyse durchzuführen und die wichtigen Textbereiche (Wörter, Teilsätze, Sätze und Abschnitte) zu identifizieren, wenden heute praktisch alle Systeme ein Ensemble von Methoden an. Jede Methode errechnet Score für jeden Textbereich. Die Scores der verschiedenen Module werden dann zu den finalen Scores kombiniert. Das System wählt dann die n Textbereiche, mit den höchsten finalen Scores aus. Das n leitet sich aus der gewünschten Summary Länge ab.

Die Performance-Score der Methoden für die Textbereich Identifikation wird normalerweise mittels der Präzision und Ausbeute ermittelt. Dafür wird je eine Textzusammenfassung von einem Menschen und dem zu bewertenden System erstellt. Die Performance-Score zeigt auf, wie nahe das Extrakt des Systems an das Extrakt des Menschen kommt. Dabei werden folgende Zahlen ermittelt: *korrekt* = Anzahl Sätze, welche vom Menschen und vom System ermittelt wurden. *falsch* = Anzahl Sätze, welche vom System ermittelt wurden, aber nicht vom Menschen. *verfehlt* = Anzahl Sätze, welche vom Menschen ermittelt wurden, jedoch nicht vom System. Mit diesen Zahlen können die Ausbeute und die Präzision errechnet werden:

$Präzision = \frac{korrekt}{korrekt+falsch}$

$Ausbeute = \frac{korrekt}{korrekt+verfehlt}$

Die Präzision gibt also an, wie viele extrahierte Sätze gut waren, wohingegen die Ausbeute angibt, wie viele gute Sätze vom System verfehlt wurden. Folgende Methoden werden für die Score Berechnung verwendet:

Positiononal Criteria: Text ist normalerweise strukturiert. Gewisse Bereiche eines Textes (Titel, Textsatz, erster Paragraph, usw.) beinhalten normalerweise wichtige Information. Für gewisse Texte (Bsp. Zeitungsartikel, wissenschaftliche Texte und technische Artikel) erzielt man sogar sehr gute Resultate, wenn man lediglich den ersten Paragraphen extrahiert und als Summary präsentiert. Mit dieser Methode kann man ein Scoring von bis zu 33% erreichen.

Cue Phrase Indicator Criteria: Für gewisse Texte zeigen Schlüsselwörter, wie z.Bsp. 'signifikant' oder 'in diesem Paper zeigen wir', dass es sich um einen wichtigen Teil handelt. Solche Sätze können mit hoher Zuverlässigkeit für die Erstellung des Summarys extrahiert werden. Zu beachten gilt jedoch, dass jeder dieser Sätze sowohl positive oder negative Attribute haben, was es zu berücksichtigen gilt.

Word and Phrase Frequency Criteria: Bereits 1959 entwickelte Luhn eine Methode, die Wichtigkeit von Wörter mittels dem Zipfschen Gesetz zu bestimmen. Das Zipfsche Gesetzt sagt aus, dass wenige Wörter sehr oft vorkommen, ein paar Wörter kommen ab und zu vor und viele Wörter kommen nur sporadisch vor. Luhn (1959) entwickelte daraus den folgenden Extraktions-Satz: *Falls ein Text gewisse Wörter unüblich oft verwendet, dann sind die Sätze, in denen diese Wörter vorkommen, wahrscheinlich wichtig.* Falls diese Methode richtig angewendet wird, erreicht man einen Score zwischen 15% und 35%. Da diese Methode für alle Sprachen relativ gut funktioniert, ist sie ideal dafür geeignet, für *Multi Lingual Summaries* eingesetzt zu werden.

Query and Title Overlap Criteria: Eine relativ effektive und einfache Methode besteht darin, die Anzahl erwünschter Wörter in einem Text zu zählen. Erwünschte Wörter sind solche, welche in Titel oder als Stichworte vorkommen.

Cohesive or lexical connectedness criteria: Wörter sind in den unterschiedlichsten Arten miteinander verbunden (Bsp.: Wiederholungen, Koreferenzen, Synonyme, usw.) Textteile können dann aufgrund ihrer Verbundenheit beurteilt werden. Sätze und Textteile, welche eine höhere Verbundenheit aufweisen, sind tendenziell wichtiger. Mit dieser Methode erreicht man Scores zwischen 30 und bis zu 60%.

Combination of Various Module Scores: In keinem der Fälle ist es gelungen, dass eine Methode gleich gute Scores wie ein Mensch erreichte. Jedoch kann durch die Kombination von verschiedenen Methoden die Score signifikant gesteigert werden. Hier wäre denkbar, die Auswahl der Methoden und die Kombination der verschiedenen Scores mittels Machine Learning durchzuführen (Bsp.: Random Forest).

Phase 2: Transformation, Interpretation oder Topic Fusion

Die erste Phase ist eigentlich das, was wir unter *Text Extraction* verstehen und auch relativ gut gelöst werden kann mit heutigen Systemen. Durch die Interpretations-Phase wird nun der extrahierte Text in ein *Abstract* umgewandelt. Hierfür werden die extrahierten Teilsätze, Sätze und Paragraphen zusammengefügt, Ausdrücken verändert und durch neue Sätze formuliert. Damit dies effizient funktioniert, braucht das System Wissen zum Kontext der Zusammenfassung. Genau dies macht es aber sehr schwierig für automatische Systeme. Bis heute gibt es keinen universell einsetzbaren Text Summarizer.

Phase 3: Synthese oder Summary generation

In der letzten Phase wird nun, das Summary aus Phase 2, welches lediglich in maschinenlesbarer Form vorhanden ist, wieder in zusammenhängende Sätze umgewandelt. Es gibt hier verschiedene Ansätze: Im Dokument Hirst u. a. (1997) wird beschrieben, wie direkt mittels 'Smoothing' ein Extrakt aus Phase 1 in eine lesbare Form gebracht werden kann. Hierfür werden (Teil-)Satzwiederholungen und Wiederholungen von Named Entities eliminiert. Ein anderer, interessanter Ansatz verfolgt Knight und Marcu (2000): Hier wird das Ergebnis der Phase 2, bei welchem die Sätze in Form eines syntaktischen Baums vorliegen, komprimiert. Dies geschieht mit einem System, welches mit dem Expectation-Maximization-Algorithmus (EM-Algorithmus) trainiert wurde. Das Ziel ist, einzelne Sätze zu kürzen. U.U. können sogar zwei Sätze zu einem oder drei Sätze zu zwei bzw. einem zusammengefasst werden.

Evaluation von Summaries

Spärck Jones und Galliers (1996) definieren 2 Arten von Summary Evaluationen: intrinsische Evaluation, bei der nur die Qualität des Outputs betrachtet wird und extrinsische Evaluation, bei der der Umfang der Userinteraktion betrachtet wird. Die meisten Evaluationsmethoden berücksichtigen nur die intrinsische Evaluation. Typischerweise wird für jeden Text ein oder mehrere Summaries von Menschen erstellt. Die Texte der zu vergleichenden Systeme werden dann mit diesen verglichen. Dabei wird die Ausbeute und die Präzision der übereinstimmenden Sätze bzw. Wörter gemessen (siehe Kapitel 'Phase 1: Analyse oder Topic Identification').

Grundsätzlich muss jedoch beachtet werden, dass es nicht *die* korrekte Zusammenfassung gibt. Daher werden in gewissen Evaluationen mehrere von Menschen erstellte Summaries für den Vergleich erstellt. Für den finalen Score wird dann der Schnitt über alle Referenzdokumente ermittelt.

Eine andere Methode, zur intrinsischen Evaluation ist, die von den Systemen erstellten Summaries von Menschen bewerten zu lassen. (Siehe Brandow und Rau 1999)

Messmethoden

Das Hauptproblem beim Evaluieren von Summaries besteht darin, dass es schwierig ist zu beschreiben, was eine gute Zusammenfassung ausmacht und wie dies gemessen werden kann. Generell kann man sich jedoch an folgende Eckpfeiler halten:

- Ein Zusammenfassung muss kürzer sein als das Original
- Die wichtigen Informationen (was jedoch Kontextabhängig ist) müssen in der Zusammenfassung vorhanden sein.

Somit können folgende zwei Messgrössen definiert werden, die aussagen, inwiefern ein Summary S mit dem original Text T übereinstimmt:

Kompressionsrate: *CR = (Länge S) / (Länge T)*

Retentionsrate: *RR = (Info in S) / (Info in T)*

Eine gute Zusammenfassung ist folglich also eine, bei welcher die Kompressionsrate so klein wie möglich und gleichzeitig die Retentionsrate so gross wie möglich ist.

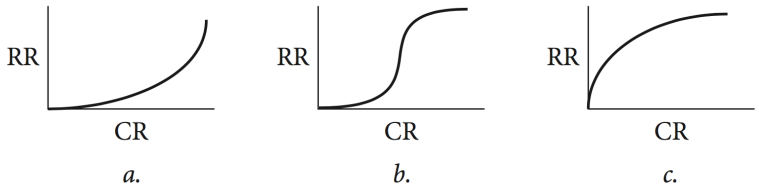

Figure 2: Kompressionsrate vs. Retentionsrate (Hovy 2005, p. 593)

Idealerweise steigt die Retentionsrate mit zunehmender Textlänge (siehe Abbild 2a). Ab einem gewissen Punkt ist diese jedoch gesättigt (siehe Abbildung 2b). Abbildung 2c zeigt, wie das Verhältnis sich bei weiter steigender Textlänge verhält.

Schlussfolgerung

Text Summarization ist ein sehr spannendes Thema, das in der jüngsten Zeit durch die massive Zunahme von frei zugänglichen Texten ein enormes Potential aufweist. Für bestimmte Bereiche wie z. Bsp. News wurde das Problem einigermassen gelöst. Trotz intensiver Forschung existiert aber bis heute immer noch kein System zur generellen Text Zusammenfassungen. Im Gegensatz dazu funktioniert Text Extraction sehr gut. Hier erreicht man durch statistische Auswertungen relativ gut Resultate. Neuere Ansätze verwenden dazu auch Machine Learning bzw. Deep Learning wie z. Bsp. RNNs.

Ein funktionierendes System würde jedoch vielfältige Probleme lösen und die künstliche Intelligenz auf ein neues Niveau heben. Es bleibt auf alle Fälle spannend und es lohnt sich, dieses Thema in Zukunft weiter zu verfolgen.

Referenzen

Brandow, K., R. und L. Rau. 1999. Automatic condensation of electronic publishing publications by sentence selection.

Choon-Ching, Ng und Ali Selamat. 2014. Text Summarization Review.

Hirst, G., C. DiMarco, E.H. Hovy und K. Parsons. 1997. Authoring and generating health-edu- cation documents that are tailored to the needs of the individual patient.

Hovy, Eduarrd. 2005. Text Summarization: 583–598.

Knight, K. und D. Marcu. 2000. Statistics-based summarization—step one: sentence compression.

Luhn, H.P. 1959. The automatic creation of literature abstracts.

Spärck Jones, K. und J.R. Galliers. 1996. Evaluating Natural Language Processing Systems: An Analysis and Review.