Text Summarization - Ein Rück- und Ausblick

Marcel Canclini, marcel.canclini@gmail.com

Juni 2017

Einleitung

Durch den Computer selbständig erstellte Zusammenfassungen von Text, sei dies aus einem oder aus mehreren Dokumenten, ist ein noch ungelöstes Machine Learning Problem. Dabei liegt es nicht daran, dass nicht geforscht wird, wie in Abbildung 1 zu sehen ist. Die Schwierigkeit liegt vielmehr darin, dass viele einzelne Problemstellungen zusammenkommen wie im ersten Kapitel aufgezeigt wird. Qualitativ hochwertige, automatische Zusammenfassungen wären in der aktuellen Zeit von Informationsüberflutung ein willkommenes Geschenk. Nicht nur die Länge von Dokumenten, auch die vielen verschiedenen Quellen mit unterschiedlichem Mehrwert machen das Auffinden und Verarbeiten von Informationen zu einer schwierigen und zeitintensiven Angelegenheit.

Der Anwendungsbereich ist immens! Sei dies in der Forschung um sich einen Überblick zu verschaffen ohne gleich jedes relevante Paper im Detail zu lesen, oder in der Medizin um schnell eine Übersicht der bereits erfolgten Behandlungen eines Patienten zu bekommen. Dies ist sicher auch ein Grund warum nicht nur die Forschung in diesem Bereich tätig ist, sondern auch grosse Firmen wie Salesforce, welche sich für ihre CRM Systeme natürlich einiges an Nutzen durch maschinell erfolgte Zusammenfassung eines Kunden erhoffen.

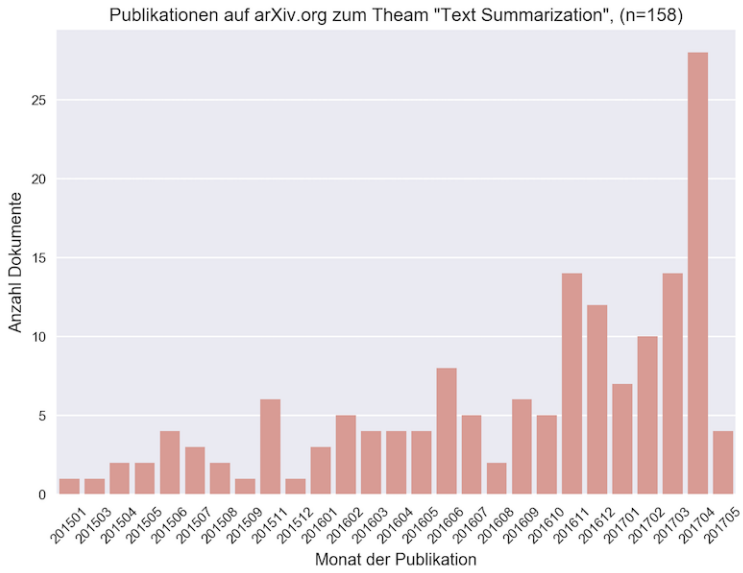

Figure 1: ungefähre Anzahl eingereichte Papers im Bereich “Text Summarization” in den letzten rund 2.5 Jahren

Problemstellungen

Warum können Computer Hunde von Katzenbildern unterscheiden oder lernen ein realistisches Gesicht zu zeichnen, aber keine Zusammenfassung eines Textes zu erstellen? Es scheint als wäre Text schwieriger maschinell zu verarbeiten als Bilder. Folgend 3 Punkte, welche aus meiner Sicht eine Erklärung dafür geben.

Abstrakt vs. Extrakt

Erstellt eine Person eine Zusammenfassung eines Textes, so erstellt er ein Abstrakt, einen neuen Text, welcher den Originaltext in einer kurzen Art wiedergibt. Ein Computerprogramm hingegen sucht sich die *wichtigsten* Sätze oder Wörter und gibt diese wieder. Also ein Extrakt aus dem Originaltext.

Qualität - Im Auge des Betrachters

Was ist eine gute, oder noch besser, die richtige Zusammenfassung eines Textes? Diese Frage ist leider nicht so einfach zu beantworten, da jede Person eine andere Ansicht davon hat. Gibt man einer Anzahl von Personen (Annotatoren) die Aufgabe ein Text zusammenzufassen, bekommt man sehr viele unterschiedliche Versionen. Das inter-annotator Agreement ist also entsprechend tief. Um einen supervised Learning Ansatz zu verwenden oder die Qualität eines Systems zu bestimmen, braucht es aber einen Gold-Standard, auch *ground truth* genannt, von Zusammenfassungen für Texte.

Die Gruppe AIPHES an der TU Darmstadt beschäftigt sich unter Anderem mit diesem Thema und erstellt anhand eines geführten Prozesses mit 7 Schritten kohärente Extrakte von deutschen Texten und stellt diese als Korpus zur Verfügung (Benikova u. a. 2016). Mit ihrem Prozess konnten sie ein inter-annotator Agreement von ca. 85% erreichen.

Die Gruppe beschäftig sich dabei nicht nur mit dem Erstellen von manuellen Zusammenfassungen, sondern wendet den erwähnten Prozess auch gleich auf deutschsprachigen Texten an. Eine auf englischem Text fuktionierende Zusammenfassung muss nämlich nicht zwingend gute Resultate in anderen Sprachen liefern. Sie schaffen somit die Basis um Modelle für deutsche Dokumente zu trainieren.

Semantic

Reines Anwenden von Machine Learning im Bereich der Textanalyse basiert fast ausschliesslich auf statistischen Merkmalen. Semantisches und syntaktisches Verständnis von Sprache wäre aber eine grosse Hilfe bei der Erstellung von verständlichen und aussagekräftigen Zusammenfassungen. Diese Tatsache erschwert wiederum die Anwendung eines einzelnen Algorithmus auf verschiedenen Sprachen.

Prozess und Ansätze

In Automatic Summarising: Factors and Directions (Jones 1998) wird der grundlegende Prozess zur Generierung von Zusammenfassungen wie folgt beschrieben:

1. source text *interpretation* to source text representation
2. source representation *transformation* to summary text representation
3. summary text *generation* from summary representation

Sparck Jones beschreibt diesen Prozess als offensichtlich. Zu beachten sind auch die klaren Datentypen, welche an der Schnittstelle der Prozessschritte verwendet werden. Die Verwendung eines solchen generischen Ansatzes erlaubt es nun verschiedene Ansätze miteinander zu vergleichen und den einzelnen Schritten zuzuordnen.

In der Vergangenheit wurden unter Anderem folgende Ansätze verwendet: (Es handelt sich hier um eine nicht vollständige Auswahl.)

Ein speziell bei News Artikeln schwer zu übertreffendes Vorgehen ist ganz einfach die ersten X Wörter eines Textes zu verwenden. Dies wird häufig auch für andere Modelle als **Baseline** genutzt.

Lexical Chaining verbindet Textfragmente oder Sätze innerhalb eines Dokumentes über Nomen, welche einen lexikalischen Zusammenhang haben. Dieser Ansatz kann nur auf einzelne Dokumente angewendet werden, ist aber ein wichtiger Baustein, auf welchem weitere Ansätze aufbauen.

Maximal Marginal Relevance wählt Sätze aus, welche eine hohe Ähnlichkeit zu bereits gesehenen Sätzen haben. Dies werden als wichtig erachtet. In einer zweiten Phase werden die redundanten Sätze wieder entfernt um eine repräsentatives Extrakt zu bekommen.

LexRank (Erkan und Radev 2004) ist eine graphenbasierte Variante inspiriert von PageRank. Hierbei wird mittels eines **unsupervised** Ansatzes unter Verwendung des lexikalischen Zentralitätsmasses ein Graph erstellt. LexRank kann auch zur Keyphrase Extraction verwendet werden.

Natürlich gibt es auch verschiedene **supervised Machine Learning** Ansätze. Hierzu braucht es aber für die Trainingsdaten manuell erstellte, mögliche Extrakte. Die Machine Learning Algorithmen versuchen nun durch Wahl der richtigen Features diese Extrakte nachzustellen.

Neuronale Netze

Natürlich dürfen auch im Bereich von Text Summarization die neuronalen Netze nicht fehlen. Diese haben in den letzten Jahren in vielen Bereichen des maschinellen Lernens gleiche oder bessere Resultate wie z.B. AlexNet (Krizhevsky, Sutskever und Hinton 2012), geliefert wobei eine Verschiebung von Feature Engineering zu Architecture Engineering erfolgt ist. Es war eine Frage der Zeit bis entsprechende Architekturen von Netzen zur Erstellung von Zusammenfassungen auftauchten.

Google: Text Summarization mit Tensorflow

2016 hat Google (Liu und Pan 2016) einen Ansatz vorgestellt, welcher mittels sequence-to-sequence learning das Erstellen von Zusammenfassungen aus kürzeren Texten automatisch lernen kann. Eine sequence-to-sequence Architektur basiert auf zwei Recurrent Neural Networks, welche Sequenzinformationen verarbeiten kann indem der erste Teil den Text liest (Encoder) und der Zweite die Zusammenfassung erstellt (Decoder).

Google sieht das Erstellen von Zusammenfassung als interessannte Aufgabenstellung für Computer, da für eine gute Zusammenfassung ein Verständnis für den Text vorhanden sein muss um die relevanten Informationen zu identifizieren. Dies wird umso schwieriger je grösser der Text wird. Google hat dazu Tensorflow Code veröffentlicht, welcher Schlagzeilen für News Artikel erstellt. Ein Blick auf die Beispiele zeigt, dass dieses Modell in der Lage ist ein Abstrakt des Textes zu erstellen und nicht nur einen *wichtigen* Satz stellvertretend für den Text als Überschrift setzt.

Salesforce: Deep Reinforcement Learning

Im Mai diesen Jahres wurde auch von Salesforce eine Deep Learning Architektur zur Erstellung von Zusammenfassungen veröffentlicht (Paulus, Xiong und Socher 2017a). Auch dieses System soll in der Lage sein einen Abstrakt des Textes zu erstellen. Gegenüber Google, soll aber der Ansatz von Salesforce nicht nur Überschriften, sondern auch längere, zusammenhängende Texte erstellen können.

Wie bei Google wird auch hier eine sequence-to-sequence Architektur verwendet. Zusätzlich wird beim Trainieren, im Stil von Reinforcement Learning, eine Feedback Schlaufe eingebaut (siehe Abbildung 2). Der *Scorer* beurteilt die gesamte Zusammenfassung und liefert einen *reward* an das Model zurück.

Figure 2: Feedbackloop im Reinforcement Prozess. Der Scorer bewertet nicht nur einzelne Wörter, sondern die gesamte Zusammenfassung (Paulus, Xiong und Socher 2017b Fig. 7)

Da auch die *ground truth* nur eine Interpretation der entsprechenden Annotatoren ist, kann nicht eine exakte Übereinstimmung für den Feedback Loop verwendet werden. Salesforce evaluiert das Resultat mittels ROGUE (Recall-Oriented Understudy for Gisting) (Lin 2004). Dabei werden einzelne Teile der generierten Zusammenfassung mit der Referenzzusammenfassung verglichen. Es muss somit nicht eine perfekte Übereinstimmung vorhanden sein. Je höher aber diese ist, desto höher ist auch der ROGUE Score. Dies wird mittels des oben erwähnten Reinforcement Prozess an das Model zurückgemeldet.

Schlussfolgerung

Die von manchen als der *heilige Gral* des Machine Learning bezeichnete Text Summarization ist auch heute nur in einfachen Bereichen wie der Zusammenfassung von News Artikeln wirklich brauchbar. Auch wenn mit aktuellen neuronalen Netzarchitekturen erstaunliche Resultate erzielt werden, braucht es noch etwas mehr um im täglichen Gebrauch zum Einsatz zu kommen.

Auch der Bericht über das von Salesforce vorgestellte System kommt zum Schluss, dass zwar bessere Resultate erzielt wurden als mit existierenden Systemen, aber qualitativ noch Lücken bestehen. Es braucht unter Anderem bessere Metriken zur automatischen Evaluation der Resultate. Eine gute Metrik muss mit den durch die Annotatoren erstellten Zusammenfassungen im Bezug auf Zusammenhang und Lesbarkeit korrelieren.

Auch wenn die Resultate in naher Zukunft sicher noch deutlich verbessert werden, wird dies nur auf englischen Texten Verwendung finden. Zusammenfassungen für andere Sprachen und über verschiedene Sprachen hinweg ist danach noch eine weitere Hürde welche zu nehmen ist.

Literatur

Benikova, Darina, Margot Mieskes, Christian M. Meyer und Iryna Gurevych. 2016. Bridging the Gap between Extractive and Abstractive Summaries: Creation and Evaluation of Coherent Extracts from Heterogeneous Sources (Dezember): 1039–1050.

Erkan, Günes und Dragomir R. Radev. 2004. LexRank: Graph-Based Lexical Centrality As Salience in Text Summarization. *J. Artif. Int. Res.* 22, Nr. 1 (Dezember): 457–479.

Jones, Karen Sparck. 1998. Automatic Summarising: Factors and Directions. In: *Advances in Automatic Text Summarization*, 1–12. MIT Press.

Krizhevsky, Alex, Ilya Sutskever und Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems 25*, hg. von F. Pereira, C. J. C. Burges, L. Bottou, und K. Q. Weinberger, 1097–1105. Curran Associates, Inc.

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In:, 74–81. Association for Computational Linguistics. (zugegriffen: 26. Juni 2017).

Liu, Peter und Xing Pan. 2016. Text Summarization with TensorFlow. *Google Research Blog*. August. (zugegriffen: 25. Juni 2017).

Paulus, Romain, Caiming Xiong und Richard Socher. 2017a. A Deep Reinforced Model for Abstractive Summarization. *arXiv:1705.04304 [cs]* (Mai). http://arxiv.org/abs/1705.04304.

---. 2017b. Your Tl;Dr by an Ai: A Deep Reinforced Model for Abstractive Summarization. https://metamind.io/research/your-tldr-by-an-ai-a-deep-reinforced-model-for-abstractive-summarization, Mai. (zugegriffen: 18. Juni 2017).