Projektbericht: Automatische Textzusammenfassung mit der Strong Nuclearity Hypothesis

Lea Wetzke
Sebastiano Gigliobianco
Daniela Weiß
Philine Huß
Department Linguistik
BSc Computerlinguistik
Diskursparsing SoSe 2022

1 Einleitung

Die automatische Erstellung von Zusammenfassungen ist in Anbetracht vieler Anwendungen nützlich. Bringt aber auch einige Schwierigkeiten mit sich. Schon alleine die Frage, was eine gute Zusammenfassung ausmacht lässt sich so ohne Weiteres nicht leicht beantworten.

Dennoch möchten wir in unserem Projekt ausprobieren, ob es möglich ist, basierend auf der "Strong Nuclearity Hypothesis" (Marcu, 2000) für Texte Zusammenfassungen automatisch zu erstellen. Unsere Forschungsfrage lautet: Ist es möglich, aus Bäumen, die mittels Rhetorical Structure Theory (RST) für jeweils einen Text generiert wurden, gute Zusammenfassungen zu generieren?

2 Theoretisches und Arbeitshypothese

Die Strong Nuclearity Hypothesis (SNH) baut auf die Idee der Rhetorical Structure Theory (RST) auf.
Die RST beschäftigt sich mit der Planung kohärenter Texte und der Analyse der globalen Struktur von Texten. Der Text wird anhand von Beziehungen, die zwischen den einzelnen Teilen (EDUs) bestehen aufgebaut, somit wird die Kohärenz von Texten durch die Erstellung einer hierarchischen Struktur zwischen den Textteilen erklärt. Die einzelnen Teile eines Textes oder auch Elementary Discourse Units (EDUs) können entweder Satelliten oder Nuklei sein. Der Nukleus beinhaltet die Kernaussage und der Satellit fügt weitere Informationen als Unterstützung hinzu. Satelliten tragen also zur Information des Nukleus bei und funktionieren oft ohne den Nukleus nicht mehr als sinnvolle Aussage. Beide Teile gehen eine Relation miteinander ein, die bestimmt auf welche Art der Satellit den Nukleus unterstützt (Hintergrundinformationen, Evidenz, Elaboration, ...). Es gibt aber auch multinukleare Relationen, in denen beide EDUs gleichermaßen bedeutungstragend für den Text sind (Kontrast, Auflistung, ...).

Laut (Marcu, 2000) sind die wichtigsten EDUs eines Textes die, welche beginnend an der Wurzel des RST-Baums der Linie des Nukleus nach unten folgend, bis hin zu einem Blatt (also einer EDU), gefunden werden. Bei Relationen, die mehrere Nuklei aufweisen, werden alle Wege durch den Baum zu den jeweiligen Blättern verfolgt. So sollen die Kernaussagen eines Textes extrahiert werden können. Wenn wir davon ausgehen, dass die Kernaussagen eines Textes eine gute Zusammenfassung des Textes darstellen, dann müssten diese EDUs eine gute Zusammenfassung darstellen. Denn die Menge der EDUs, die durch den Algorithmus der Strong Nuclearity Hypothesis extrahiert werden, stellen die Essenz des Textes dar.

Unsere Arbeitshypothese: die Zusammenfassungen, die mit der Strong Nuclearity-Hypothese generiert werden, sind aussagekräftige Zusammenfassungen - das heißt, sie sind besser als randomisiert erstellte oder baseline-Zusammenfassungen.

3 Daten

Unsere grundlegende Datenquelle ist das Potsdam Commentary Corpus (PCC) (Stede, 2004), bestehend aus 176 deutschen Zeitungskommentaren aus der Märkischen Allgemeinen Zeitung (MAZ). Die Korpustexte sind bereits mit den für dieses Projekt benötigten Annotationen wie RST-Bäume, grammatische Rollen und Koreferenzketten versehen.

4 Vorgehen

Im Rahmen dieses Projektes vergleichen wir nach der SNH automatisch generierte Zusammenfassungen aus dem Potsdam Commentary Corpus (PCC) mit Baseline- und randomisiert generierten, sowie neuronalen Zusammenfassungen. Als Goldstandard werden handannotierte Zusammenfassungen von PCC-Kommentaren genutzt.

4.1 Gold-Zusammenfassung

Die Gold-Zusammenfassungen wurden von menschlichen Annotatoren erstellt. Dazu wurden für jeden Kommentar 3 vollständige Sätze markiert, welche die beste Zusammenfassung des Textes ausmachen sollen.

4.2 Random- und Baseline-Zusammenfassungen

Auch für die Baseline- und Random-Zusammenfassungen haben wir uns darauf geeinigt bei diesem Schema zu bleiben.
Für die automatisch erstellten Random-Zusammenfassungen werden zufällig drei ganze Sätze aus dem jeweiligen Kommentar gezogen, wobei die Reihenfolge der Sätze beliebig sein kann.
Für die ebenfalls automatisch generierten Baseline-Zusammenfassungen werden jeweils der erste, der mittlere und der letzte Satz ausgewählt.

4.3 SNH-Zusammenfassung

Um die SNH-Zusammenfassungen automatisch zu erstellen, haben wir die RST-Bäume als NLTK.ParentedTree eingelesen und rekursiv durchlaufen, um nach der SNH die Kernaussagen zu extrahieren.
Innerhalb des PCCs sind immer auch die Überschriften als Nukleus markiert, welcher immer auch als Kernaussage mit extrahiert wird. Da in den Gold-Zusammenfassungen die Überschriften nicht inkludiert sind, haben wir uns dafür entschieden auch die Überschriften aus den SNH-Zusammenfassungen auszuschließen.
Die Ergebnisse des SNH-Algorithmus sind eine unstete Anzahl an EDUs: mal ist es nur eine EDU, mal auch deutlich mehr als drei. Und die extrahierten EDUs sind nicht immer ganze Sätze. Das macht die Evaluation im Endeffekt schwierig, da dies von der Version der handannotierten Gold-Zusammenfassungen als auch von der Anzahl in den Baseline- und Random-Zusammenfassungen abweicht.[1]

4.4 neuronale Zusammenfassungen

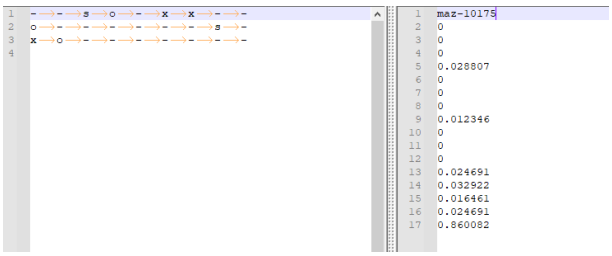
Als weiteres Vergleichsmittel haben wir außerdem neuronale Zusammenfassungen erstellen, einmal mit 25000 und einmal 50000 Trainingsschritten. Für die Erstellung der neuronalen Zusammenfassungen haben wir ein Encoder-Decoder Transformer Modell (Vaswani et al., 2017) trainiert mit Texten sowohl aus dem Korpus der German Text Summarization Challenge (Cieliebak et al., 2019) als auch aus dem MLSUM-Datensatz (Scialom et al., 2020). Das Transformer-Modell hat Shared Embeddings, Encoder und Decoder benutzen also die selben Embeddings für den Input. Der Encoder liest die Texte, und der Decoder generiert die Zusammenfassungen. Zum Trainieren des Modells wurde kaggle benutzt.[2]

5 Evaluation

5.1 Entity-Based Coherence Modell

Ursprünglich war vorgesehen, die verschiedenen Zusammenfassungen mittels des Entity-Based Coherence Modells nach Barzilay and Lapata (2008) zu evaluieren. Dazu wurden aus den mmax-Dateien des PCC die annotierten grammatischen Rollen (S, O, X, -) und die Koreferenzketten der Entitäten für jeden Kommentar extrahiert und ein Entity Grid (siehe Figure 1) erstellt, aus denen die Transitions-Wahrscheinlichkeiten der Permutationen aller vier grammatischen Rollen berechnet wurden, dargestellt in einem Vektor (vgl. (Barzilay and Lapata, 2008) S.8 §3.3).

Figure 1: Entity Grid und Vektor von maz-10175



[1] Die genaue Implementierung kann in unserem Repository auf GitLab eingesehen werden: https://gitup.uni-potsdam.de/gigliobianco/rst/-/tree/main

[2] https://www.kaggle.com/code/sebag90/german-zusammenfassung-ml-rst

Mit diesen Textvektoren wird dann eine Support Vector Machine (SVM) trainiert, um die Qualität der Zusammenfassungen zu evaluieren. Für das Training werden immer zwei Zusammenfassungen durch die SVM miteinander verglichen. Für jeden Text gehen wir davon aus, dass die Gold-Zusammenfassungen besser als die Baseline sind, die wiederum besser als die Random-Zusammenfassungen sind. Mit dieser Annahme wird auch die SVM trainiert: als Input dienen Paare aus baseline/random, gold/baseline, gold/random, wobei jeweils das erste immer besser als das zweite ist (vgl. Ranking-Problem Barzilay and Lapata (2008), S.11, §4.1). Multipliziert man dann die Textvektoren der Zusammenfassung mit der Gewichtsmatrix der trainierten SVM, erhält man einen Qualitätsscore der Zusammenfassung.

Für die Evaluation sollten die Vektoren der Strong-Nuclearity-Zusammenfassungen mit der Gewichtsmatrix der SVM multipliziert, um die Qualität zu bewerten und um diese mit den anderen Zusammenfassungen zu vergleichen. Unser erwartetes Ergebnis war, dass die SNH-Zusammenfassung besser ausfällt als die des Random- und Baseline-Algorithmus.

Bereits beim Training der SVM stellte sich jedoch heraus, dass der Entity-Based Coherence-Ansatz mit unseren Texten kein zufriedenstellendes Ergebnis lieferte. Die Textvektoren konnten der SVM allem Anschein nach nicht als Grundlage für eine zuverlässige Klassifizierung dienen. Wir vermuten, dass dies an den sehr kurzen Texten liegt. Denn damit einher geht ebenfalls eine geringe Entitätenanzahl und damit auch wenige Transitions der grammatischen Rollen. So kam es öfter vor, dass mehrere Texte die gleiche Anzahl und Art von Transitions aufwiesen und daher genau gleiche Vektoren erhalten. Beim Vergleich durch die SVM von zwei identischen Vektoren kann also keiner der beiden Vektoren als "besser" klassifiziert werden. Auch Fälle, in denen unsere Zusammenfassungen gar keine Entitäten enthalten, können mit diesem Ansatz nicht ausgewertet werden.
Hinzu kommt: damit die SVM sowohl einzelne Texte als auch Paare betrachten kann, werden die Paare als Subtraktion zweier Vektoren dargestellt: $P(V_1, V_2) = V_1 - V_2$ wenn $Label_P = 1$. Wenn wir sehr dünnbesetzte Vektoren haben, z.B. [0.9, 0, 0, 0, 0.1] und [0, 0, 1, 0, 1], sind die Paar-Vektoren sehr ähnlich: [0.9, 0, -1, 0, 1] bzw. identisch, wenn von einem Vektor ein 0-Vektor subtrahiert wird. Daher waren die Ergebnisse der Klassifizierung nicht viel besser als die mit einer zufällig generierten Gewichtsmatrix der SVM.

5.2 Evaluation durch Satzvektoren mit RoBERTa und der SVM

Aus diesem Grund erfolgt die Evaluierung durch Textvektoren aus einem RoBERTa-Modell (Liu et al., 2019). Dieser Ansatz hat außerdem den Vorteil, dass wir damit auch neuronale Zusammenfassungen in unseren Vergleich mit aufnehmen konnten.

RoBERTa berechnet Vektoren für die Sätze unserer verschiedenen Zusammenfassungen. Anschließend wir die SVM trainiert, um die Textqualität zu beurteilen. Dabei erhalten Paare $P(t_i, t_j)$ das Label $LP = 1$, falls t_i besser als t_j ist, bzw. $LP = 0$, falls t_i schlechter als t_j ist.
Vektor-Paare werden folgendermaßen berechnet: $P(t_i, t_j) = t_i - t_j$.

Dabei nehmen wir an:

$$gold > baseline > random$$

Die Klassifizierung erfolgt nun mithilfe der SVM. Für jede Kommentar-Zusammenfassung wird ein Score ermittelt, indem wir das Skalarprodukt zwischen dem Gewichtsvektor der SVM und dem Vektor der Zusammenfassung berechnen.

6 Probleme

6.1 Fehlerhafte Annotationen

Bei Stichproben einiger Kommentare stellten wir fest, dass nicht alle Annotationen zuverlässig waren. So stießen wir beispielsweise auf falsch annotierte Koreferenzen, wie beispielsweise bei maz-10110. An manchen Stellen sind Satzteile als Entitäten markiert, die unserer Meinung nach eigentlich keine wären. Dies könnte unser Ergebnis etwas verzerrt haben, da wir nicht jede einzelne Datei auf Richtigkeit prüfen konnten.

6.2 Überschriften

Teilweise wurden bei der SNH-Zusammenfassung auch die Überschriften der Kommentare als eine der besten EDUs mitgezählt, während in den anderen Zusammenfassungen (auch bei den handannotierten Gold-Zusammenfassungen) die Überschrift nicht berücksichtigt wird. Unsere Lösung hierfür ist, Überschriften nicht in den SNH-Zusammenfassungen zu erlauben.

6.3 Satzindices

Ein weiteres Problem waren zunächst auch die unterschiedlichen Satzindices. Nicht nur unsere Baseline- und Random-Summaries (satztokenisiert mit NLTK) erhielten andere Satzindices als die handannotierten Gold-Kommentare, sondern auch die Annotationen der mmax-Dateien aus dem PCC unterschieden sich sowohl von den Gold-Indices als auch von unseren. Zur Vereinheitlichung der Indices haben wir die Tokens abgeglichen und die entsprechenden Entitäten aus den mmax-Dateien herausgesucht.

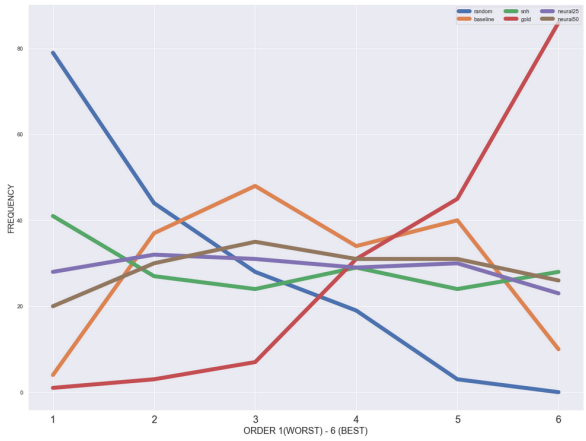
6.4 Probleme bei neuronalen Zusammenfassungen

Diese Zusammenfassungen aus den Trainingsdaten sind sehr kurz (meistens 1 Satz) und sehr oft wird der erste Satz als Zusammenfassung angesehen. Das Modell lernt also einfach den ersten bzw. zweiten Satz aus dem Text zu extrahieren. Dieser wird jedoch noch modifiziert, sodass eine Auswertung mit dem Entity Grid nicht erfolgen könnte. Daher war eine Auswertung nur mit einem Transformer-Modell möglich.

7 Ergebnisse

Im Diagramm in Figure 2 ist die Auswertung der verschiedenen Zusammenfassungsarten abgebildet. Auf der X-Achse ist die Order abgebildet, also die Ränge der Zusammenfassungen auf einer Skala: 1 ist sehr schlecht, 6 ist sehr gut. Auf der Y-Achse ist die Frequenz, wie oft eine Zusammenfassung eines bestimmten Typs diesen Rang erhielt.
Random (blau) ist der schlechte Vergleichswert: diese Zusammenfassungen haben am meisten die Order 1. Im Mittelfeld befindet sich, wie erwartet, baseline (orange), diese Zusammenfassungen sind am meisten in der Order 3. Die Gold-

Figure 2: Ergebnisse als Diagramm



Zusammenfassungen (rot) sind am meisten auf dem ersten Platz.
Die neuronalen (lila für 25k und braun für 50k) und die SNH-Zusammenfassungen (grün) liegen alle im Mittelfeld und sind miteinander vergleichbar, was die Qualität der Zusammenfassungen angeht. Alle drei Zusammenfassungsarten liegen etwas unter baseline, bei allen drei ist die Verteilung der Ränge relativ gleichmäßig.
Die SNH-Zusammenfassungen sind also nicht extrem gut: sie sind nicht nah am Goldstandard und ihre Qualität ist leicht unter der der baseline-Zusammenfassungen. Dies kann man auch weiterführend an Figure 3 erkennen: hier sind konkrete Zahlen aufgelistet, die aufzeigen, wie oft SNH besser als eine andere Zusammenfassungsmethode war. Hier ist baseline mit 92 besseren Zusammenfassungen als SNH etwas stärker.

Figure 3: Ergebnisübersicht

```
snh besser als gold: 38
snh besser als baseline: 81
snh besser als neural25: 85
snh besser als neural50: 80
snh besser als random: 114
```

```
gold besser als snh: 135
baseline besser als snh: 92
neural25 besser als snh: 88
neural50 besser als snh: 93
random besser als snh: 59
```

8 Zusammenfassung

Im Rahmen des Projektes haben wir untersucht, ob Zusammenfassungen, welche mithilfe der Strong Nuclearity Hypothesis generiert wurden, prägnant sind. Dazu haben wir andere Zusammenfassungsmethoden als Vergleichswerte herangezogen und zuerst versucht, die Performanz aller Zusammenfassungen mittels Entity Coherence zu evaluieren. Da dies aufgrund der

Kürze der Zusammenfassungen zu Problemen geführt hat, haben wir eine Alternative für die Evaluierung gewählt in Form von Textvektoren aus einem RoBERTa-Modell, mit welchen wir eine Support Vector Machine trainiert haben. In unseren Ergebnissen ließ sich feststellen, dass Zusammenfassungen nach der SNH nur sehr mittelmäßig sind: Wenn wir also zurück auf unsere Arbeitshypothese schauen, dann haben wir diese widerlegt. Mithilfe der SNH (so wie wir sie hier implementiert haben und für die kurzen Kommentare aus dem PCC) lassen sich nicht prägnante, sondern nur passable Zusammenfassungen generieren. Es wäre interessant zu schauen, ob SNH-Zusammenfassungen bei längeren Texte und/oder für Texte in englisch bessere Ergebnisse liefern.

9 Ausblick

Ist eine Zusammenfassung, die der Bewertung des Entity-Based Coherence Modell nach sehr kohärent ist, wirklich eine präzise Zusammenfassung für einen Kommentar? Was macht eine gute Zusammenfassung aus? Auf welche Kriterien muss hierbei geachtet werden, neben der Kohärenz? Eine Möglichkeit, um wirklich zu erkennen, was eine gute Zusammenfassung ist, wäre nochmal eine menschliche Bewertung der Zusammenfassungen durchzuführen. Denn: wie akkurat sind eigentlich all diese Bewertungsmechanismen? Man könnte Studie durchführen, in welcher Proband*Innen die Qualität von Zusammenfassungen bewerten. Oder aber auch eine andere Studie, in der menschlich bewertet wird, wie gut die Bewertungen sind und somit mehr Klarheit darüber finden, was eine gute Zusammenfassung eigentlich ausmacht.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Mark Cieliebak, Don Tuggener, and Fernando Benites, editors. 2019. *Proceedings of the 4th Swiss Text Analytics Conference, SwissText 2019, Winterthur, Switzerland, June 18-19, 2019*, volume 2458 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. CogNet.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

M. Stede. 2018. *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*. narr studienbücher.

Manfred Stede. 2004. The Potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.