Informationsextraktion durch gezielte Zusammenfassung von Texten

Timm Euler*

*Universität Dortmund, Fachbereich Informatik, Lehrstuhl für Künstliche Intelligenz, Email: euler@ls8.cs.uni-dortmund.de

Abstract. Heutige automatische Textzusammenfassung erstellt meistens Extrakte, also ausgewählte Teile des Originales, als Zusammenfassung. Fast alle Arbeiten auf diesem Gebiet befassen sich mit generischen Extrakten, die den Text als Ganzes wiedergeben sollen. Im Gegensatz dazu werden hier *gezielte* Extrakte, die nach bestimmten Inhalten suchen, vorgestellt als Alternative zur klassischen Informationsextraktion. Durch die Verwendung von maschinellem Lernen beschränkt sich die Vorgabe von Inhalten auf die Markierung von Dokumenten und Dokumentteilen als relevant oder nicht. Eine zweite Stufe entfernt unwichtige Satzteile aus den extrahierten Textsegmenten. Eine Anwendung zur Verkürzung von Emails zu SMS-Nachrichten wurde mit guten Ergebnissen implementiert.

Keywords. Information Extraction, Text Summarization, Sentence Compaction

1 Einleitung

Eine von Menschen erstellte Zusammenfassung eines Textes ist meistens eine neuformulierte Wiedergabe der für wesentlich gehaltenen Inhalte eines Textes (engl. *abstract*). Dies ist sehr schwierig zu automatisieren, da dazu u.a. tiefes Textverständnis, breites Weltwissen und eine gute Spracherzeugung nötig sind. Solche Ansätze existieren als Theorie (etwa [16]), doch beschränken sich heutige automatische Verfahren meist auf *extracts*, also die Wiedergabe von ausgewählten, unveränderten Ausschnitten aus dem Ausgangstext ([4]). Hierauf konzentriert sich auch dieser Artikel.

In beiden Fällen kann man nach dem Zweck der Zusammenfassung unterscheiden: *generische* Zusammenfassungen versuchen, möglichst alle wesentlichen Aspekte des Ausgangstextes unterzubringen und so dem Leser einen allgemeinen Eindruck des längeren Textes zu geben. Zur *gezielten* Zusammenfassung (*tailored summary*, [5]) berücksichtigt man dagegen nur bestimmte Inhalte, die im Benutzerinteresse liegen. Dafür muss eine inhaltliche Spezifikation vorliegen. Im Falle der klassischen Informationsextraktion wird diese in Form von genauen Templates und Slots gegeben, zusammen mit aufwendigen Beschreibungen, wie das richtige Template gewählt wird und wie die Slotinformation im Text gefunden werden kann. Hier gibt es Ansätze, mit maschinellem Lernen den Aufwand zu reduzieren (eine gute Übersicht findet sich in [3]). Eine einfachere Art der inhaltlichen Vorgabe stellt eine Liste von Stichwörtern dar, die ebenfalls mit maschinellem Lernen gewonnen werden kann, wie dieser Artikel zeigen wird. Die gezielte Wiedergabe nur der gewünschten Inhalte stellt eine andere Art von Informationsextraktion dar.

Im folgenden werden in Abschnitt 2 Arbeiten zur Zusammenfassung mittels Extraktion aufgeführt und einige Schwierigkeiten diskutiert. Danach folgt eine Beschreibung des eigenen Ansatzes (Abschnitt 3). Abschnitt 4 erläutert eine Anwendung, die bestimmte Emails mit Hilfe gezielter Zusammenfassung sowie der Kürzung der ausgewählten Sätze zu SMS-Nachrichten[1] verkürzt, und gibt Resultate der Experimente und Evalutionen dazu.

2 Erstellung von Extrakten

Zur Erstellung von Textextrakten gehen die meisten Verfahren satzweise vor und entscheiden, ob ein bestimmter Satz zum Extrakt gehören soll oder nicht ([4]). Deshalb wird dieses Vorgehen auch als Satzfiltern bezeichnet. Statt Sätzen werden in [11] und [14] Absätze als kleinste Einheiten verwendet. Die Entscheidung zur Extraktion kann getroffen werden aufgrund der Position im Dokument, aufgrund bestimmter Redewendungen (*cue phrases*, z.B. *Zusammenfassend lässt sich sagen, ...*), nach statistischen Maßen wie Worthäufigkeiten, oder nach semantischer Ver-

[1] SMS = Short Message Service, Kurznachrichten für Mobiltelefone

wandtschaft von Wörtern (*lexical cohesion*). Für eine Übersicht siehe [5].
Die Lesbarkeit der Extrakte leidet unter dem fehlenden Zusammenhang der einzelnen Sätze. Insbesondere ist bei Pronomen oft nicht mehr klar, worauf sie sich beziehen (Anaphora-Problem). In der Regel soll daher der Extrakt nicht den Ausgangstext komplett ersetzen, sondern nur deutlich machen, ob es sich lohnt, diesen zu lesen.
Ein Problem bei generischen Extrakten ist die Evaluation ([7]). Wenn einige Menschen gebeten werden, die allgemein relevantesten Sätze eines Textes auszuwählen, erzielen sie oft weniger als 50% Übereinstimmung ([5]). Ist also der Zweck der Zusammenfassung unbekannt, so scheint ein Vergleich zweier Extrakte aus einem Text schwierig. Man kann aber die Verwendbarkeit des Extraktes für bestimmte Aufgaben untersuchen, zum Beispiel Textklassifikation oder Relevanzeinordnung (vgl. [7] oder [6]).
Mit inhaltlichen Vorgaben kann das Satzfiltern leichter werden, wenn etwa ganze Textteile außer Acht gelassen werden können. Ebenso wird die Bewertung von gezielten Zusammenfassungen erleichtert: es lässt sich nachprüfen, ob die Fragen des Benutzers bei Kenntnis des Extraktes beantwortbar sind, ob also alle gewünschten Informationen aus der inhaltlichen Vorgabe im Extrakt auftauchen. Solche Bewertungen standen bei der Informationsextraktion immer im Vordergrund, wurden jedoch für gezielte Zusammenfassungen selten durchgeführt ([6]). Arbeiten, die gezielte Zusammenfassungen behandeln, sind [10] und [13].

3 Gezielte Satzextraktion

Als inhaltliche Vorgabe zum gezielten Satzfiltern verwendet das hier vorzustellende Verfahren Stichwortlisten. Diese Listen erstellt es automatisch. Zwar gibt es zur automatischen Erkennung von Index- oder Schlüsselwörtern bereits viel Literatur (etwa [1]; eine gute Übersicht ist in [15] zu finden); dort wird aber versucht, die jeweils wichtigsten Wörter für einen ganzen Text oder eine Textsammlung zu finden. Dagegen kommt es hier darauf an, nur Wörter zum vorgegebenen Thema zu sammeln. Im folgenden Abschnitt wird das Vorgehen dafür beschrieben. Darauf folgt die Beschreibung der Satzextraktion mit Hilfe dieser Stichwortliste (Abschnitt 3.2).

3.1 Ermittlung der Stichwortliste

In einem Trainingskorpus (Sammlung von Texten) werden alle Sätze, die das gewünschte Thema behandeln, markiert. So erhält man eine Menge von Beispielsätzen, aus denen gelernt werden kann. Das Ergebnis des Lernens ist eine Rangliste von Stichwörtern, denen jeweils ein Gewicht zugeordnet ist.
Vor dem Lernen der Gewichtung ist es sinnvoll, eine sogenannte Stammformenreduktion durchzuführen. Dabei werden verschiedene Formen desselben Wortes (für das Wort `laufen` z.B. `lief`, `läufst` usw.) normalisiert auf dieselbe Stammform (etwa `lauf`). Man benötigt dazu ein Lexikon der verschiedenen Wortformen. Ohne Stammformenreduktion wurden bei dem hier vorgestellten Verfahren keine guten Ergebnisse erzielt.
Für die Gewichtung jedes Wortes, das in markierten Sätzen (positiven Beispielen) vorkommt, wurden verschiedene Verfahren getestet, die nun kurz vorgestellt werden.

Worthäufigkeit Die einfachste Möglichkeit ist, die (absolute) Anzahl der Vorkommen eines Wortes in den positiven Beispielsätzen durch die Anzahl seiner Vorkommen in den anderen Sätzen zu dividieren und so ein Gewicht zu erhalten. Leicht verfeinernd, kann man die relative Häufigkeit pro positivem Satz durch die relative Häufigkeit pro negativem Satz teilen.

Information Gain Dieses Maß wird zum Beispiel in [12] erläutert. Es beruht auf dem grundlegenden Maß der Entropie. Jedes Wort erhält hierbei sein Gewicht nach seiner Unterscheidungskraft zwischen positiven und negativen Beispielen (Sätzen).

G^2-Statistik G^2 ([1]) ist ein statistisches Maß, das angibt, zu welchem Grad die Häufigkeit eines Wortes in einem Text (hier in einem Satz) höher ist, als es nach seiner Gesamthäufigkeit im Korpus und der Länge seines Textes (Satzes) zu erwarten wäre. Hier wird die Berechnung abgewandelt und statt dessen gemessen, wieviel höher die Häufigkeit eines Wortes in den positiven Beispielsätzen ist, als es nach der Häufigkeit in den negativen Beispielen zu erwarten wäre, die also als repräsentativ für beliebige Texte gelten.

SVM-Gewichte Support Vector Machines (SVMs) werden erfolgreich zur Textklassifikation eingesetzt ([8]). Deshalb wurde hier versucht, sie zur Klassifikation von Sätzen einzusetzen, was aber fehlschlug (F-Maß 16%, zu F-Maß siehe Abschnitt 3.2). Dennoch kann das Ergebnis einer SVM-Rechnung eingesetzt werden. Für die Klassifikation von Texten mit SVMs wählt man eine Repräsentation der Texte durch Vektoren: jede Stelle des Vektors entspricht einem möglichen Wort. Der Wert des Vektors an dieser Stelle ist im einfachsten Fall 0 oder 1, je nachdem, ob das Wort in dem Text vorkommt; für das hier vorgestellte Verfahren wurde der Wert durch das *Tf-idf*-Maß des Wortes bestimmt (siehe z.B. [5]). Die Trainingsbeispiele (Texte bzw. Sätze) liegen also als je ein Vektor in einem mehrdimensionalen Vektorraum vor. Darin versucht

Verfahren	Recall	Precision	F-Maß	Fallout	Recall	Precision	F-Maß	Fallout
Worthäufigkeit	83.5	79.2	81.2	2.3	82.9	71.7	76.5	3.4
Information Gain	70.5	69.3	69.6	3.6	80.4	71.8	75.6	4.6
G^2-Statistik	73.3	71.5	72.3	3.1	74.9	76.8	75.7	4.7
SVM-Gewichte	79.9	71.2	75.1	5.1	57.5	85.5	68.5	1.1

Tabelle 1 Ergebnisse des gezielten Satzfilterns in Prozent, je nach Verfahren zur Erzeugung einer Rangliste. Im rechten Teil die Werte bei dokumentweiser Erstellung der Stichwortlisten.

die SVM dann, eine Hyperebene so zu legen, dass die Beispiele richtig in positiv und negativ getrennt werden und zusätzlich der Abstand zu den nächstgelegenen Beispielen möglichst großist. Die Koeffizienten der Gleichung für diese Hyperebene kann man auffassen als Gewicht, mit dem jedes Wort für die Lage der Ebene berücksichtigt wurde. Es ergibt sich also aus der Ebenengleichung direkt eine Gewichtung für jedes Wort. Eine deutliche Verbesserung der Stichwortliste bzw. der mit ihr erzielten Ergebnisse wurde erzielt, indem nur SVM-Gewichte von Wörtern mit einem hohen *Tf-idf*-Gesamtwert verwendet wurden.
Nach Anwendung eines dieser Verfahren steht eine Rangliste von gewichteten Wörtern zur Verfügung, mit der man nun beliebige Sätze klassifizieren kann, wie im nächsten Abschnitt beschrieben ist.

3.2 Satzauswahl

Für neue, ungesehene Sätze muss zunächst wieder die Stammformenreduktion durchgeführt werden. Anschließend wird jedem Wort des Satzes sein Gewicht aus der Rangliste zugeordnet (wenn es dort vorkommt), die Gewichte werden addiert und die Summe durch die Anzahl der Wörter im Satz geteilt, um längere Sätze nicht zu bevorzugen. Damit hat jeder Satz ein eindeutiges Gewicht.
Auf dem Trainingskorpus kann nun ein Schwellwert bestimmt werden, über dem ein Satzgewicht liegen muss, damit der Satz positiv klassifiziert wird (in die Zusammenfassung aufgenommen wird). Ein hoher Schwellwert lässt nur wenige Sätze zu, die aber mit höherer Sicherheit tatsächlich zum Thema gehören. Ein niedriger Schwellwert klassifiziert mehr Sätze positiv, verpasst also weniger tatsächlich relevante Sätze, lässt aber mit höherer Wahrscheinlichkeit einige nichtrelevante Sätze zu. Dies wird in den folgenden beiden Standardmaßen ausgedrückt: *Recall* ist der Anteil an den tatsächlich themenrelevanten Sätzen, die das Verfahren positiv klassifiziert. *Precision* ist der Anteil an den positiv klassifizierten Sätzen, die tatsächlich relevant für das gewünschte Thema sind. Diese Maße sind nicht unabhängig voneinander, wie die Diskussion des Schwellwertes deutlich macht.

Um den besten Schwellwert für einen Textkorpus zu finden, werden auf der Trainingsmenge mehrere Schwellwerte ausprobiert und jeweils Recall (R) und Precision (P) gemessen. Um einen direkten Vergleich zu ermöglichen, kann man das sogenannte *F-Maß* bilden (siehe z.B. [9]), das sich wie folgt berechnet:

$$F = \frac{2RP}{R+P}\,.$$

4 Anwendung und Experimente

Die oben beschriebene gezielte Textzusammenfassung ermöglicht vielfältige Anwendungen. Zu Testzwecken wurde ein Email-to-SMS-Service simuliert, bei dem Benutzer die für sie wichtigen Emails auf ihr Mobiltelefon weiterleiten lassen können, wenn sie gerade nicht am Arbeitsplatz sind. Für dieses Experiment gelten alle Emails als wichtig, die mit Terminabsprachen zu tun haben, die also vielleicht sogar den Tagesablauf des Benutzers beeinflussen können. Es sollen dabei nur die terminbezogenen Sätze aus der Email extrahiert, danach gekürzt und weitergeleitet werden. Dazu zählen Ankündigungen oder Terminverschiebungen ebenso wie Ab- oder Zusagen.
Es wurde ein Korpus von 560 deutschsprachigen Emails (über 45000 Wörter) verwendet, von denen die Hälfte Terminabsprachen beinhaltet, aber nicht unbedingt ausschließlich Terminabsprachen. Die andere Hälfte besteht aus beliebigen Emails. Etwa 13% der Sätze des gesamten Korpus wurden als terminbezogen markiert. Für jeden Lauf wurden neun Zehntel des Korpus zum Training verwendet, das verbleibende Zehntel zum Testen.

4.1 Satzfiltern

Bei den Tests zum Satzfiltern, die für jedes in Abschnitt 3.1 beschriebene Verfahren durchgeführt wurden, ergaben sich die Recall- und Precisionwerte aus Tabelle 1 (linker Teil), die zehnfach kreuzvalidiert sind (zum Fallout-Wert siehe unten). Die Standardabweichung der Werte für Recall, Precision und F-Maß über die zehn Runden liegt jeweils zwischen 4 und 7 Prozent, die für Fallout unter 1 Prozent.

Die besten Ergebnisse werden demnach erzielt vom einfachen Verfahren des Zählens der Worthäufigkeiten und vom eher komplexeren Verfahren der Verwendung von Gewichten, die von einer Support Vector Machine ermittelt wurden. Aber auch die beiden anderen Verfahren liefern brauchbare Resultate. Zusätzlich hat man die Möglichkeit, durch Senkung des Schwellwertes den Recall zu verbessern, weil man vielleicht ungern terminbezogene Informationen verpassen möchte; dann sinkt allerdings die Precision.
Die Precision-Werte sind abhängig von der Zusammensetzung des Korpus: Wenn mehr Emails ohne Terminbezug dazukommen, werden auch mehr Sätze ohne Terminbezug positiv klassifiziert. Ergänzend zur Precision wird daher der *Fallout*-Wert angegeben, also der Anteil der negativen Sätze, die irrtümlich positiv klassifiziert werden ([9]). Wie man sieht, ist dieser Wert recht niedrig, man kann daher auch für realistischere Korpora annehmen, dass die Anzahl der irrtümlich positiv klassifizierten Sätze in einem vertretbaren Rahmen bleibt.
Ein weiteres Experiment zeigt, dass die Ergebnisse im selben Bereich liegen, wenn man die Markierung der Beispiele nicht satzweise, sondern textweise vornimmt (rechter Teil von Tabelle 1). Es muss also nur jede Email als terminbezogen oder nicht markiert werden, was sicherlich weniger Arbeit erfordert, als jeden Satz zu behandeln. Die Gewinnung der Stichwortlisten funktioniert dann genauso.

4.2 Erstellung der SMS

Heutige SMS-Nachrichten sind standardisiert auf eine Länge von 160 Zeichen begrenzt. Für einen nicht geringen Anteil der Emails genügt das auch nach der automatischen Satzfilterung nicht. Um mehr Informationen in einer SMS unterbringen zu können, wurde ein Verfahren zur Kürzung von Sätzen erprobt. Dieses beruht auf der Erkennung von Wortarten mit Hilfe eines Lexikons sowie von grammatischen Teilphrasen mit endlichen Automaten. Dazu wurde das sprachverarbeitende System MESON eingesetzt, ein Nachfolger der am Deutschen Forschungsinstitut für Künstliche Intelligenz entwickelten sprachverarbeitenden Teile des SMES-Systems ([2]); MESON liefert auch die Stammformenreduktion.
Die Grundidee besteht darin, die erstellte Stichwortliste als Hinweis auf die wichtigen Stellen eines Satzes zu verwenden. Dabei werden je nach nötigem Kürzungsfaktor verschiedene Stufen der Satzkürzung eingesetzt. Auf den niedrigen Stufen werden übliche Abkürzungen eingeführt, zum Beispiel `FR` statt `Freitag`; es werden Anrede- und Grußformeln gestrichen und meistens inhaltsleere Wörter wie `naja` oder `überhaupt` entfernt. Danach werden Wörter mit Wortarten, die meistens für das Verständnis des Satzes nicht erforderlich sind, entfernt, nämlich Artikel und Adjektive/Adverben, jedoch nur, wenn sie nicht in der Stichwortliste stehen. Eingeklammerte Teile eines Satzes können ebenfalls gestrichen werden. Zuletzt werden ganze Phrasen entfernt, allerdings keine Verbalphrasen, da Verben die syntaktische Struktur eines Satzes bestimmen und für das Verständnis dringend erforderlich sind. Phrasen werden aber nur entfernt, wenn sie kein Wort aus der Stichwortliste enthalten. Dies soll dafür sorgen, dass die Satzteile, die für die Auswahl des Satzes gesorgt haben, also relevant für das vorgegebene Thema sind, erhalten bleiben und die entsprechende Information in der SMS erkennbar ist.
Bei Anwendung der höheren Stufen kann jedoch leicht auch wichtige Information verlorengehen, indem der Satzzusammenhang entstellt wird (das Kürzungsverfahren ist fast rein syntaktisch, nur die Stichwortlisten bieten eine gewisse semantische Orientierung). Es entsteht ein Tradeoff zwischen Lesbarkeit der SMS und Unterbringung von möglichst viel Information. Erkenntnisse dazu werden im nächsten Abschnitt beschrieben. Ein Beispiel eines Emailtextes mit zwei dazu erstellten SMS-Texten, die auf unterschiedlich starker Kürzung basieren, gibt Abbildung 1. Bei der ersten SMS wurden Artikel und Adjektive, aber keine Phrasen außer der Grußformel entfernt. Die Lesbarkeit der zweiten SMS leidet unter den entfernten Phrasen, dafür ist der Starttermin enthalten. Von der Emailadresse des Absenders wird nur der erste Teil (vor dem @) in die SMS übernommen, vom Betreff nur die ersten 20 Zeichen. Das Zeichen ^ steht für ein oder mehrere entfernte Wörter, # steht für weggelassene Sätze.

4.3 Evaluation

Die Bewertung des Satzfilterns wurde schon in Abschnitt 4.1 vorgestellt. Wie in Abschnitt 2 erläutert wurde, kann bei gezielten Zusammenfassungen aber auch die Erhaltung der interessierenden Information leichter gemessen werden. Zum Terminkomplex können zum Beispiel konkrete Fragen zu Zeit oder Ort des Termins gestellt werden. Der Email-to-SMS-Service wurde mit einem Fragebogen bewertet, der solche Fragen stellt und der Testpersonen vorgelegt wurde, die entweder die ursprüngliche Email kannten oder nur eine SMS mit gefilterten und evtl. gekürzten Sätzen. Das Augenmerk galt dabei der Beantwortbarkeit der Fragen, nicht der Korrektheit der Antworten, die wohl nur vom Absender der Email bestimmt werden könnte. Die Ergebnisse lassen sich wie folgt zusammenfassen:

- Die Information, die am besten erhalten bleibt, ist der Zeitpunkt des Termins. Auch sein Status, also ob er ausfällt oder verschoben wird etc., ist

```
From: kupferstecher@noel.cs.uni-freiland.de
Betreff: Nächstes Treffen und Pacman Demo
Liebe A4lerinnen und A4ler, das nächste Treffen findet am 18.10.2000 im
Raum Campus Süd, GB IV, R. 110 um 10 Uhr statt. Hiermit mchte ich um
Vorschläge für die Tagesordnung bitten. Vorher würde ich gerne die
gwünschte Demonstration von Pacman durchführen, sofern Herr Zeisig und
Beatrix vorher schon Zeit haben 9 Uhr 30 als Starttermin für die Demo
sollte ausreichen. Viele Grüße, Maria

kupferstecher(Nächstes Treffen und): ^Treffen findet am 18.10.2000 im Raum
Campus Süd, GB IV, R. 110 um 10 Uhr statt.#^würde ich^gwünschte
Demonstration von Pac

kupferstecher(Nächstes Treffen und): ^Treffen findet am 18.10.2000 im^IV,^.
110 um 10 Uhr statt.#^würde^durchführen, sofern^haben 9 Uhr 30 als
Starttermin^sollt
```

Abbildung 1 Eine Termin-Email und zwei dazu erstellte SMS.

meistens noch aus den verkürzten Texten erkennbar. Was für ein Termin es ist, kann leichter verloren gehen, ließe sich aber für den Adressaten der Email oft leicht aus dem Kontext erschließen.

- Wenn Informationen zu Ort des Termins oder beteiligten Personen im Ursprungstext enthalten sind, gehen diese oft verloren. Der Grund ist, dass die Wörter dafür nicht nur in terminbezogenen Kontexten auftauchen und daher nicht in der Stichwortliste enthalten sind, oder nur mit geringem Gewicht. Dementsprechend werden solche Satzteile gekürzt oder die entsprechenden Sätze weggefiltert.
- Die Verständlichkeit einer SMS litt auch unter dem fehlenden Hintergrundwissen der Testpersonen, wie viele bemerkten: Wären sie selbst Absender oder Empfänger der Email gewesen, so hätten sie mehr Angaben machen können.
- Bei starker Kürzung wird der sich ergebende Text oft unverständlich.

Wegen des letzten Punktes erscheint es besser, auf einige Informationen in der SMS zu verzichten und dafür die Teile, die untergebracht werden können, verständlich zu halten. Bei mittlerer Kürzung, die nicht mehr die Phrasenentfernung durchführt, sollte in den meisten Fällen Halt gemacht werden. Dann kann die Zusammenfassung in vielen Fällen die Lektüre des Originals ersetzen.

5 Zusammenfassung und Ausblick

Es wurde ein Verfahren zur gezielten Satzextraktion vorgestellt, das als inhaltliche Vorgabe nur eine Markierung interessierender Sätze verlangt. Es genügt sogar zur Stichwortgewinnung, die Markierung dokumentweise vorzunehmen. Beim Satzfiltern können damit gut 80% Performanz nach dem F-Maß gewonnen werden, zumindest in der Termindomäne. Durch eine gezielte Auswertung des Informationsgehaltes der Extrakte wurde deutlich, welche Informationen am besten erkannt werden. Das Verhältnis von Recall und Precision ist über einen einzigen Schwellwert leicht anpassbar an die Anwendung. Trainingszeiten zum Lernen der Stichwortlisten liegen im Bereich weniger Stunden im Falle der Verwendung einer SVM; bei Benutzung der Worthäufigkeiten ist die Laufzeit vernachlässigbar.
Die zweite Stufe der Kürzung, in der unwichtige Teile aus Sätzen entfernt werden, sollte moderat eingesetzt werden, kann jedoch die Dichte der Informationspräsentation erhöhen. Bei zu starker Kürzung sinkt die Lesbarkeit zu sehr.
Dieses Verfahren muss noch für andere Domänen und andere Sprachen getestet werden. Weiterhin bietet es sich an, es für mehr als eine Domäne parallel anzuwenden. Dazu würde man Texte verwenden, die in mehrere Klassen eingeteilt wurden, und für jede Klasse eine Stichwortliste ermitteln. Alle Sätze, die mindestens einer Stichwortsammlung genügen, würden schließlich extrahiert. Alternativ kann man vielleicht sogar mit den beiden Klassen *interessant* und *uninteressant* auskommen, um eine einzige Stichwortliste zu erhalten. Diese wäre dann auch bei Verschiebung der Interessen des Benutzers leicht nachlernbar. Man würde allerdings eine größere Menge Daten benötigen, als bei den obigen Experimenten verwendet wurde, da für jedes Unterthema einige Stichwörter in der Liste enthalten sein müssen.

6 Danksagung

Dieser Artikel ist die Zusammenfassung meiner Diplomarbeit am Lehrstuhl für Künstliche Intelligenz der Universität Dortmund. Ich bedanke mich bei Prof. Dr. Katharina Morik und Dipl.-Inform. Ralf Klinkenberg für die ausgezeichnete Betreuung.

Literatur

1. J.D. Cohen. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, 46(3), 1995. Man beachte auch das Erratum in Bd. 47, 3, S.260.
2. Thierry Declerck, Judith Klein, and Guenter Neumann. Evaluation of the nlp components of an information extraction system for german. In *Proceedings of the first international Conference on Language Resources and Evaluation (LREC) 1998*, pages 293–297, Granada, 1998.
3. Oren Glickman and Rosie Jones. Examining machine learning for adaptable end-to-end information extraction systems. In *Proceedings of the AAAI 1999 Workshop on Machine Learning for Information Extraction*, 1999.
4. Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of ACM-SIGIR'99*, pages 121–128, 1999.
5. Udo Hahn and Inderjeet Mani. Automatic text summarization. Tutorial for the Fifteenth National Conference on Artificial Intelligence (AAAI), Madison, Wisconsin, July 1998.
6. T. Firmin Hand and B. Sundheim, editors. *TIPSTER-SUMMAC Summarization Evaluation. Proceedings of the TIPSTER Text Phase III Workshop*, Washington, 1998.
7. H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods: Experiments and analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, pages 60–68. AAAI, 1998.
8. T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137 – 142, Berlin, 1998. Springer.
9. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of SIGIR 95*, pages 246–254, 1995.
10. Inderjeet Mani and Eric Bloedorn. Machine learning of generic and user-focused summarization. In *Proceedings of AAAI'98*, Madison, Wisconsin, Juli 1998.
11. M. Mitra, A. Singhal, and C. Buckley. Automatic text summarization by paragraph extraction. In I. Mani and M. Maybury, editors, *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spanien, Juli 1997.
12. John Ross Quinlan. *C4.5: Programs for Machine Learning*. Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.
13. Ellen Riloff. A corpus-based approach to domain-specific text summarisation: A proposal. In B. Endres-Niggemeyer, J. Hobbs, and K. Sparck-Jones, editors, *Workshop on Summarising Text for Intelligent Communication*, Dagstuhl, BRD, 1993.
14. T. Strzalkowski, J. Wang, and B. Wise. A robust practical text summarization system. In *AAAI Intelligent Text Summarization Workshop*, pages 26–30, Stanford, CA, März 1998.
15. Peter D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000.
16. T.A. van Dijk. Semantic macro-structures and knowledge frames in discourse comprehension. In M.A. Just and P.A. Carpenter, editors, *Cognitive Processes in Comprehension*, pages 3–32. Lawrence Erlbaum, Hillsdale, NJ, 1977.