

Exposé zur Studienarbeit

Textklassifikation von Erdbebenmeldungen



Institut für Informatik
Humboldt-Universität zu Berlin

Mario Lehmann

16. Mai 2011

Betreuer: Prof. Dr. Ulf Leser, Lars Döhling

1 Motivation

Die Suche nach Möglichkeiten der automatisierten semantischen Analyse von geschriebener natürlicher Sprache ist im Zuge der immer weiter fortschreitenden Informationsflut, der wir ausgesetzt sind, unabdingbar. Durch immer größer werdende Datenmengen und den begrenzten Fähigkeiten des Menschen diese vollständig aufzunehmen und zu verarbeiten, sind immer mehr vorgeschaltete Selektions- und Filterungsprozesse notwendig. Diese sorgen für eine Reduktion des Suchraums und ermöglichen es somit dem Menschen, wieder Herr über die riesigen Datenmengen zu werden [FD92].

Ein Beispiel dafür ist die automatisierte Analyse von Nachrichtendaten [JR90]. Insbesondere in einem Katastrophenfall müssen die helfenden Stellen sehr schnell die notwendigen Informationen zur Verfügung haben, um zeitnah und effizient reagieren zu können. Um die Arbeit zu erleichtern, könnten dabei zukünftig automatisierte Systeme eingesetzt werden, die aus dem Nachrichtenstrom diejenigen Informationen herausfiltern, die für die Katastrophe irrelevant sind. Aufgrund der reduzierten Informationsmenge wäre es dadurch für die helfenden Stellen möglich, die tatsächlich relevanten Informationen in viel geringerer Zeit zu finden.

Einen ersten Ansatz für ein solches Verfahren bietet das EquatorNLP-Projekt von Lars Döhling [Döh10]. Forschungsschwerpunkt war dabei die automatisierte Extraktion von Informationen über Erdbebenschäden aus Nachrichtentexten. Dabei konnten nur Nachrichtentexte verwendet werden, die über Erdbeben handelten, also domänenspezifisch waren. In der folgenden Arbeit soll Equator-NLP daher um einen vorgeschalteten Filter erweitert werden, der es zulässt, aus beliebigen Nachrichtentexten jene auszuwählen, die von Equator-NLP danach ausgewertet werden können.

2 Zielstellung

Genauer ist es Ziel, zu überprüfen, ob das Hauptthema eines Nachrichtenartikels automatisch klassifiziert werden kann. Hierbei beschränkt sich die Arbeit auf die Klassifikation von Artikeln, die das Thema „Erdbeben und ihre Auswirkungen“ behandeln. In einem weiteren Schritt soll überprüft werden, ob es auch möglich ist, die Nachrichtenartikel einem konkreten Erdbeben zuzuordnen.

3 Vorgehen

Die Klassifikation erfolgt in zwei Stufen (Abbildung 1). In der ersten Stufe sollen zunächst alle Artikel, die über Erdbebenauswirkungen berichten, von den anderen Artikeln getrennt

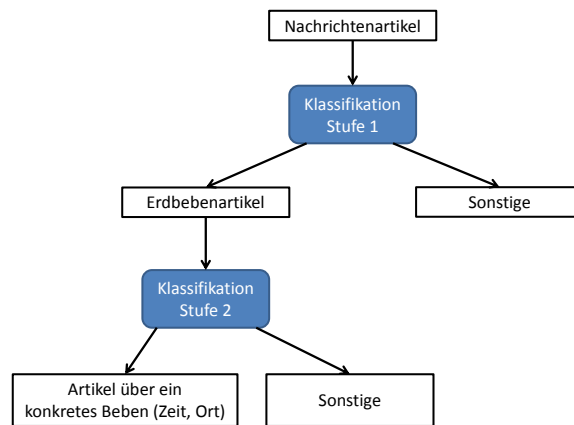


Abbildung 1: Klassifikationsschema

werden. In der Menge der Erdbebenartikel erfolgt in der zweiten Stufe die Klassifikation hinsichtlich eines konkreten Erdbebens, welches durch die beiden Parameter Zeit und Ort (als Geokoordinaten) determiniert ist. Genau wie in der ersten Stufe erhält man zwei Mengen von Artikeln, die disjunkt voneinander sind. Zum einen die Menge aller Erdbebenartikel, die über das konkrete Erdbeben berichten und zum anderen die Menge aller sonstigen Erdbebenartikel.

4 Methoden

4.1 Korpusgenerierung

Für das Anlernen der Klassifikatoren und die Evaluierung des Klassifikationsergebnisses wird ein annotiertes Korpus benötigt. Ein Korpus enthält Texte, ihre Metainformationen sowie Annotationen, also Informationen über die Textklasse bzw. über bestimmte Attribute innerhalb der Texte. Mit Hilfe dieser Annotationen werden in der Trainingsphase die Klassifikatoren angelern. In der Testphase werden dann anschließend die Annotationen derjenigen Texte aus dem Korpus, die für den Lernvorgang nicht verwendet wurden, dazu benötigt, um das Ergebnis der Klassifikation hinsichtlich seiner Güte einschätzen zu können.

In unserem Fall soll ein Korpus erstellt werden, für den Nachrichtentexte aus dem Internet extrahiert werden sollen. Die Nachrichten sollen dabei mehreren englischsprachigen Quellen entstammen. Das Korpus soll dabei auch heterogene Dokumente enthalten, also solche,

die über mehrere Naturereignisse berichten. Dies ist eine Herausforderung, da viele Artikel zwar über ein konkretes Ereignis handeln, allerdings auch noch andere Ereignisse in diesem Zusammenhang erwähnen, was die Suche nach dem Hauptthema eines Artikels vermutlich erschwert.

Worüber ein Artikel bzw. ein Teil dessen genau handelt, muss manuell annotiert werden und fließt als Information in das Korpus mit ein. Für die Klassifikation auf zweiter Stufe sollen u. a. im Text enthaltene zeitliche und räumliche Attribute ausgewertet werden. Hierfür müssen die Nachrichten des Korpus zusätzlich temporal und spatial annotiert werden. Dieser Vorgang erfolgt automatisch unter Verwendung bereits bestehender Werkzeuge. Das Annotationsergebnis dieser Werkzeuge wird a-priori als korrekt betrachtet, was ein Problem darstellen könnte, da die Programme nicht alle Attribute korrekt annotieren können. Inwieweit sich dieser Fehler auf das Klassifikationsergebnis auswirkt, wird Bestandteil der Diskussion. Für die Annotation zeitlicher Entitäten soll hierbei das Programm *HeidelTime* [[SG10a], [SG10b]] Verwendung finden. Für die spatiale Annotation bieten sich *MetaCarta*¹ oder *OpenCalais*² an.

4.2 Klassifikation

Für die Textkategorisierung stehen mittlerweile eine Vielzahl von Klassifikatoren zur Verfügung. Die Klassifikatoren verwenden dabei unterschiedliche Modelle, woraus sich unterschiedliche Methoden für den Lernvorgang ergeben. Diese Modelle basieren auf gewissen Merkmalen (*features*) des Textes, die in gewissen Merkmalsklassen unterteilt sind. *Surface Features* sind Merkmale, die sich direkt auf das Wort bzw. die bedeutungstragenden Sequenzen von Zeichen - sogenannte Token [Döh10] - in einem Text beziehen. So kann das Wort an sich schon ein Merkmal sein, aber auch sein POSTag (Wortklasse) oder die Wortlänge. Auch *Character N-grams* also die Abfolge von N Zeichen, oder diskrete Eigenschaften des Textes (Kodierung, Struktur bestimmter Token, etc.) können ein Merkmal bilden. Des Weiteren gibt es die Klasse der *Context Features*, in der analysiert wird, wie ein Wort in seinem Kontext eingebettet ist (An welcher Stelle des Textes steht das Wort? Welche Worte/Attribute treten in unmittelbarer Umgebung noch auf?). Auch abstraktere Features wie die Häufigkeit eines Wortes oder auch die häufigste Wortklasse für einen bestimmten Term kann ein Merkmal für die Textklassifikation sein. Welche Merkmale sich genau für das Klassifikationsproblem in der Arbeit anbieten (*feature selection*), soll hierbei ebenso Untersuchungsgegenstand sein wie die Wahl des geeigneten Klassifikators. In der Arbeit möchte ich dabei drei Klassifikatoren genauer untersuchen:

¹<http://www.metacarta.com>

²<http://www.opencalais.com>

Naive Bayes: Naive Bayes in der Textklassifikation [MN98] basiert darauf, dass unterschiedliche Merkmalsvorkommen in den verschiedenen Klassen existieren. Für jedes Dokument und jede Klasse wird dabei die Wahrscheinlichkeit berechnet, dass ein Dokument bzw. Vektor aus Merkmalen zu einer bestimmten Klasse gehört. Die maximale Wahrscheinlichkeit entscheidet dann schließlich die Klassenzugehörigkeit. Aus mathematischer Sicht handelt es sich bei Naive Bayes um ein Modell, das jedem Punkt im Merkmalsraum eine Klassenzugehörigkeit zuordnet, bei der das geringste Risiko einer Fehleinschätzung besteht (*minimum risk decision* [ZYL10]).

Entscheidungsbäume: Die Klassifikation mit Entscheidungsbäumen [LJ98] ist eine baumartige Abfolge von Entscheidungsregeln, wobei auf jeder Entscheidungsstufe ein konkretes Merkmal ausgewertet wird. Dabei wird sehr oft dasjenige Merkmal verwendet, das im Hinblick auf den Informationsgewinn (*information gain*) die größte Trenneigenschaft besitzt.

Support Vector Machine (SVM): Bei der SVM [Joa98] handelt es sich um ein Verfahren, bei dem im Merkmalsraum eine Trennfunktion (Hyperebene) berechnet wird, die am weitesten von den Klassengrenzen bzw. von deren speziellen Vertretern (Support-Vektoren) entfernt ist. Sind die Klassengrenzen dabei nicht linear trennbar, so kann die SVM die Trainingsvektoren bzw. den Merkmalsraum auch in einen höherdimensionalen Raum transformieren, in welchem die Klassen möglicherweise wieder linear trennbar sind. Nach der Rücktransformation des Merkmalsraums können sich dadurch Klassengrenzen von sehr komplexer Struktur ergeben, wodurch u. U. auch schwierigere Klassifikationsprobleme gelöst werden können.

Getestet werden die Klassifikatoren gegen einen Baseline-Klassifikator. Dies erfolgt über die einfache Klassifikation über das Wort „earthquake“ bzw. auf zweiter Stufe über das Erstellungsdatum des Artikels. Für die Klassifikation wird der *RapidMiner*¹ eingesetzt, da er alle erwähnten Klassifikatoren enthält und zudem über sehr gute Visualisierungsmöglichkeiten verfügt.

Es ist zu vermuten, dass in der Arbeit gerade auf der ersten Klassifikationsstufe die *Surface Features* eine wichtige Rolle spielen werden. Da bestimmte Themen auch das Vorkommen bestimmter Worte generieren, erwarte ich mir von dieser Merkmalsklasse die größten diskriminativen Effekte. Auf der zweiten Stufe werden vermutlich die kontextuellen Merkmale eine größere Relevanz besitzen, da sich die Texte aufgrund der thematischen Ähnlichkeit kaum mehr wesentlich in ihren *Surface Features* unterscheiden sollten. Da wir auf dieser Stufe nach einem konkreten Erdbeben suchen, sollen außerdem temporale und spatiale Merkmale ausgewertet werden. Hierfür könnte es erforderlich sein, zeitliche und örtliche Mengen aus

¹<http://rapid-i.com>

den Parametern des gesuchten Ereignisses zu generieren. So kann beispielsweise das Datum eines Ereignisses innerhalb eines Artikels durch einen relativen Begriff (yesterday, last week, etc.) ausgedrückt sein. [[SG10a], [SGP10]]. Für das Suchen einer konkreten Zeitangabe wäre hierbei also eine Ersetzung (Normalisierung) der relativen Begriffe mit ihren konkreten Daten bzw. Datumsintervallen notwendig. Ähnlich verhält es sich mit räumlichen Entitäten. So kann in einem Artikel ein genauer Ort, an dem das Beben stattfand, erwähnt werden oder auch sein Oberbegriff (Land, Kontinent, etc.) oder gar die nächst größere Stadt [CS07]. Für die Auflösung dieser räumlichen Entitäten bietet sich die *GeoNames*¹-Datenbank an, da sie eine große Menge geographischer Orte mit ihren Koordinaten enthält. Ausgehend von der Koordinate des gesuchten Bebens lässt sich damit eine Liste aller möglichen Ortsbezeichnungen in einem bestimmten Umkreis generieren, die ebenfalls mit dem konkreten Ereignis assoziiert sein können. Enthält also ein Artikel den gesuchten Zeitpunkt und eine Ortsangabe, die räumlich nah zum gesuchten Ort liegt, ist anzunehmen, dass die Wahrscheinlichkeit sehr hoch ist, dass er genau über das gesuchte Ereignis berichtet.

4.3 Evaluation

Die Güteeinschätzung der Klassifikation erfolgt über Kreuzvalidierung, wobei als Gütekriterium Precision, Recall und das F_1 -Maß verwendet werden soll. Da mehrere Klassifikatoren benutzt werden, bietet sich ein direkter Vergleich an. Da die Güte der Klassifikation auf zweiter Stufe von der ersten Stufe abhängt, würde ich auch gerne die zweite Stufe separat auf ihre Güte hin untersuchen. Dazu sollen nur die Erdbebenartikel aus dem Korpus betrachtet werden, die zuvor manuell als solche annotiert wurden. Da diese von vornherein über Erdbeben handeln, ist ausgeschlossen, dass bei der Suche nach einem konkreten Beben ein Artikel in die Positivklasse fällt, obwohl er gar nicht über Erdbeben handelt. Interessant wäre hierbei auch die Untersuchung des Güteunterschieds, wenn man die Klassifikatoren der zweiten Stufe mit lediglich den vorannotierten Erdbebenartikeln trainiert im Gegensatz zur Verwendung des gesamten Korpus für den Lernvorgang.

5 Verwandte Arbeiten

Mittlerweile gibt es eine Vielzahl von Arbeiten, die sich mit Textklassifikation beschäftigen. Einen guten Überblick über unterschiedliche Methoden dafür gibt „*Classification of Text Documents*“ von Li und Jain [LJ98]. Dort werden die Klassifikatoren Naive Bayes, Nearest Neighbor, Entscheidungsbäume und die Subspace Methode verwendet, um Texte von der *Yahoo! Newsgroup* zu klassifizieren und deren Ergebnis miteinander zu vergleichen. Außerdem untersuchen die Autoren Möglichkeiten der Feature Selection. Als Ergebnis weisen alle

¹<http://www.geonames.org>

Klassifikatoren eine ähnlich hohe Accuracy von ca. 70-80% auf. Lediglich der Verlauf der Accuracy in Abhängigkeit der verwendeten Featureanzahl variiert unter den Klassifikatoren. Eine Arbeit, die den gleichen Untersuchungsgegenstand wie die geplante, nämlich Nachrichtentexte, hat, ist „*Classifying News Stories using Memory Based Reasoning*“ [MLW92]. Hier wird die k-Nearest-Neighbor-Methode bzw. eine variierte Form davon dafür verwendet um Nachrichtentexte zu kodieren, also thematisch einzustufen. Als Ergebnis erhalten sie auf ein Korpus von 50000 Nachrichtentexten einen Recall von 83% bei einer Precision von 88%. Einen anderen Weg beschreiten Chan et al. in ihrer Arbeit „*Automated Online News Classification with Personalization*“ [CSL01], in welcher sie Support-Vector-Machines für die Klassifikation von Online Nachrichten verwenden. Als Trainingskorpus diente dabei die Reuters Test Collection ¹. Anschließend wurden Nachrichten eines Nachrichtenchannels (*Yahoo! News*) in regelmäßigen Abständen aus dem Web extrahiert, in eine Datenbank abgelegt, mit der trainierten SVM klassifiziert und bei entsprechender Benutzeranfrage wieder angezeigt. Außerdem bestand die Möglichkeit für den User selber Kategorien über eine Menge von Keywords anzulegen. Leider gibt die Arbeit keine konkreten Informationen über die Performance des Systems, allerdings bietet sie viele hilfreiche Informationen darüber, wie man ein Nachrichtensystem in Verbindung mit SVMs entwickelt. Zuletzt sei auch noch die Studienarbeit „*Extraktion von räumlichen und zeitlichen Informationen aus Webtexten*“ von Nora Popp [Pop07] erwähnt, in der unterschiedliche Methoden für die Identifikation von raumzeitlichen Entitäten aufgezählt und erläutert werden. Unter Anderem wird dort auch das Tool *MetaCarta* beschrieben, welches auch ich evtl. verwenden möchte.

¹<http://www.daviddlewis.com/resources/testcollections/reuters21578>

Literatur

- [CS07] N. Cardoso and M.J. Silva. Query expansion through geographical feature types. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 55–60. ACM, 2007.
- [CSL01] C.H. Chan, A. Sun, and E.P. LIM. Automated Online News Classification with Personalization. The 4th International Conference on Asian Digital Libraries, 2001.
- [Döh10] L. Döhling. Erkennen von Angaben über Erdbebenschäden in Textmeldungen. Diplomarbeit, Humboldt-Universität zu Berlin, 2010.
- [FD92] P.W. Foltz and S.T. Dumais. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60, 1992.
- [Joa98] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142. Springer-Verlag, 1998.
- [JR90] P.S. Jacobs and L.F. Rau. SCISOR: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.
- [LJ98] Y.H. Li and A.K. Jain. Classification of text documents. *The Computer Journal*, 41(8):537, 1998.
- [MLW92] B. Masand, G. Linoff, and D. Waltz. Classifying news stories using memory based reasoning. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–65. ACM, 1992.
- [MN98] A. McCallumzy and K. Nigamy. A comparison of event models for naive bayes text classification, 1998.
- [Pop07] N. Popp. Extraktion von räumlichen und zeitlichen Informationen aus Webtexten. Studienarbeit, Humboldt-Universität zu Berlin, 2007.
- [SG10a] J. Strötgen and M. Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics, 2010.
- [SG10b] J. Strötgen and M. Gertz. TimeTrails: A System for Exploring Spatio-Temporal Information in Documents. *Proceedings of the VLDB Endowment*, 3(2), 2010.

- [SGP10] J. Strötgen, M. Gertz, and P. Popov. Extraction and exploration of spatio-temporal information in documents. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, page 16. ACM, 2010.
- [ZYL10] B. Zhou, Y. Yao, and J. Luo. A Three-Way Decision Approach to Email Spam Filtering. *Advances in Artificial Intelligence*, pages 28–39, 2010.