Portierung eines Relationsextraktionssystems auf eine neue Domäne

Niko Felger

Bachelorarbeit zur Erlangung des Grades eines

Bachelor of Science in Computerlinguistik

Vorgelegt an der Universität des Saarlandes

15. August 2007

Gutachter:

Prof. Dr. Hans Uszkoreit

Dipl.-Ling. Feiyu Xu

Inhaltsverzeichnis

Danksagung

Prof. Dr. Hans Uszkoreit bin ich zu Dank verpflichtet für die eingehende Beratung im Vorfeld dieser Arbeit und für die Vermittlung des Themas. Bei Dipl.-Ling. Feiyu Xu möchte ich mich für die hervorragende Betreuung und den vorgegebenen Rahmen bedanken. Ihr Feedback und die regelmäßigen Diskussionen in Gesprächen und Emails haben einen wesentlichen Beitrag zu dieser Arbeit geleistet. Beide standen mir auch bei der Klärung organisatorischer Probleme äußerst hilfreich zur Seite.

Eine große Hilfe war auch Dipl.-Ing. Li Hong, die mir sehr dabei geholfen hat, die Einzelheiten des bestehenden Systems zu verstehen und auf meine vielen Fragen immer gute Antworten geben konnte. Außerdem möchte ich Rui Wang für die sehr aufschlussreiche Diskussion über Minipar danken; Michaela Regneri dafür, dass sie mich mit den Erfahrungen, die sie bei ihrer Bachelorarbeit gemacht hat, auf viele formale und organisatorische Punkte hingewiesen und vor einigen Fehlern bewahrt hat; meiner Mutter, Nils Reiter, Benjamin Roth und Sarah Haggenmüller, für das Korrekturlesen dieser Arbeit und die vielen hilfreichen Anmerkungen. Sarah Haggenmüller danke ich besonders auch für ihre ständige Unterstützung und ihre scheinbar endlose Geduld.

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

1 Einleitung

Informationsextraktion ist ein sprachtechnologisches Teilgebiet, dem immer mehr Aufmerksamkeit zuteil wird. Nicht zuletzt angetrieben von den enormen Fortschritten, die innerhalb des letzten Jahrzehnts sowohl bei der Named Entity–Erkennung und der Extraktion einfacher binärer Relationen, als auch bei der Entwicklung flacher Textverarbeitungswerkzeuge und Machine Learning–Techniken erzielt wurden, hat es sich zu einem bedeutsamen Forschungsthema entwickelt, mit dem große Hoffnungen verknüpft werden.

Im Gegensatz zur klassischen computerlinguistischen Herangehensweise an die semantische Analyse von Texten, die auf ein möglichst vollständiges Textverstehen abzielt, erkennen moderne Informationsextraktionsmethoden Referenzen auf bestimmte Typen von Entitäten, wie etwa Personen, Organisationen und Datumsausdrücken, und extrahieren bestimmte, im Allgemeinen im Voraus definierte Relationen, die zwischen diesen Entitäten bestehen.

Motiviert durch den Wunsch, einen Nutzen aus dem Reichtum an Information zu ziehen, wie ihn große Textsammlungen wie etwa das World Wide Web oder MEDLINE[1] bieten, wurden effiziente Methoden entwickelt, die strukturierte Bedeutungsrepräsentationen aus unstrukturiertem Text extrahieren können, ohne dabei die einzelnen Sätze vollständig „verstehen“ zu müssen.

Derartige Techniken wurden eingesetzt, um eine ganze Reihe kommerzieller und nicht-kommerzieller Anwendungen zu realisieren. Citeseer (`citeseer.ist.psu.edu`) extrahiert automatisch Informationen über Autoren, Titel, Referenzen und vieles mehr aus wissenschaftlichen Publikationen. Froogle (`froogle.google.com`) untersucht eine Vielzahl von Onlineshops und präsentiert dem Anwender Preisvergleiche für einander ähnliche Produkte. Weitere Anwendungen sind etwa der Aufbau von Datenbanken über offene Stellen, Bildungsangebote oder Versicherungsfälle. (Für weitere Beispiele siehe McCallum (2005).)

In den meisten dieser Fälle handelt es sich allerdings entweder um teilweise strukturierten Text wie HTML-Dokumente oder wissenschaftliche Artikel, oder um Fälle in denen lediglich Named Entities und binäre Relationen extrahiert werden. Diese Systeme sind außerdem üblicherweise sehr stark an die Domäne angepasst, innerhalb der sie operieren, und nur mit großem Aufwand auf neue Domänen übertragbar. *Domänenunabhängige* Verfahren, die mit *freiem Text* umgehen können und dabei *komplexe Relationen oder Ereignisse* extrahieren, sind noch ein sehr aktives Forschungsthema.

Ein Verfahren, das all diese Eigenschaften aufweist, ist in Xu u. a. (2006) und Xu u. a. (2007) beschrieben. Im Kern steckt ein Algorithmus, der völlig unabhängig von der gesuchten Relation und der Art des Textes arbeitet und Regeln auf der Basis syntaktischer Information erstellt, mit deren Hilfe

[1] `www.ncbi.nlm.nih.gov/entrez`

Instanzen der Relation gefunden werden können. Kein Verfahren ist jedoch vollständig domänenunabhängig, da eine Implementierung stets auf eine bestimmte Domäne zugeschnitten ist. Teile dieser Implementierung, wie z.B. die Named Entity-Erkennung, müssen für eine neue Domäne geändert oder angepasst werden. Man spricht daher von *domänenadaptiven Verfahren*.

In der vorliegenden Arbeit habe ich das System von Hong (2006), welches das Verfahren von Xu u. a. (2006) für die Extraktion von Instanzen des Ereignisses *Preisverleihung* in der Nobelpreisdomäne implementiert, auf die Domäne der Musikpreise portiert. In der Arbeit wird berichtet, welcher Aufwand mit einer solchen Portierung verbunden ist und welche Ergebnisse bei der Extraktion erzielt wurden. Die für die Domäne der Musikpreise gesammelten Textdaten weisen einige Eigenschaften auf, die die Extraktion gegenüber den Nobelpreisdaten erschweren (siehe Abschnitt 5). Xu u. a. (2006) erwarten, dass die Regeln, die für eine Domäne gewonnen wurden, auf eine verwandte Domäne übertragen werden können. Somit können sie bei der Extraktion von Relationen helfen, für die die automatische Regelinduktion problematisch ist, etwa weil die Instanzen dieser Relation nur selten erwähnt werden. Ich habe ein Experiment durchgeführt, in dem die Regeln aus der Extraktion der Nobelpreisverleihungen für die Extraktion der Musikpreisverleihungen angewendet wurden. Der Vergleich der Ergebnisse mit der Extraktion ohne diese Regelübertragung spricht für ein solches Vorgehen.

Der Rest der Arbeit ist wie folgt aufgebaut. In Abschnitt 2 werde ich einen Überblick über verwandte Forschungsarbeit geben, Abschnitt 3 stellt die Problemstellung und die Motivation für diese Arbeit dar, Abschnitt 4 beschreibt die Implementierung und die nötigen Anpassungen bei der Portierung auf eine neue Domäne. In Abschnitt 5 werden die Ergebnisse der Extraktion und der Regelübertragung präsentiert. Abschließend bietet Abschnitt 6 eine Zusammenfassung der Arbeit.

2 Stand der Forschung

2.1 Ansätze zur Relationsextraktion

2.1.1 Regelbasierte Systeme

Manuelle Konstruktion. Der klassische Ansatz zur Relationsextraktion ist die Anwendung von Regeln auf den zu untersuchenden Text. Diese Regeln haben als linke Seite typischerweise Muster, die direkt auf der Oberfläche des Textes arbeiten, wie etwa reguläre Ausdrücke, oder eng mit ihr verbunden sind, wie es bei syntaktischen Mustern der Fall ist. Ein Verfahren, solche Muster zu erhalten, ist deren manuelle Konstruktion basierend auf dem Wissen von Domänenexperten und Korpusuntersuchungen (z. B.: Blaschke und Valencia, 2002).

Die händische Erstellung dieser Regeln ist allerdings sehr kompliziert

und teuer bzw. aufwändig und die erzeugten Regeln sind im Allgemeinen nicht zwischen Domänen übertragbar. Hinzu kommt, dass solche Systeme häufig stärkeren Beschränkungen der erreichbaren Recall–Werte unterliegen als andere Systeme, da es äußerst schwierig ist, die große Variabilität der natürlichsprachlichen Ausdrucksmöglichkeiten für Relationen vorherzusehen.

Automatische Konstruktion. Um diese Beschränkungen zu überwinden, bietet sich die Verwendung von Machine Learning–Methoden an, um automatisch Regeln zu extrahieren. Diese Methoden werden allgemein nach dem Umfang der beim Training verfügbaren Information über korrekte und falsche Extraktionsentscheidungen in *'überwachte'* (englisch: *supervised*) und *'halb–überwachte'* (*semi–supervised*) unterschieden, wobei letztere auch als *'minimal überwachte'* (*minimally supervised*) oder *'Bootstrapping'*–Verfahren bekannt sind.

Überwachte Verfahren verwenden ein Korpus, in dem sämtliche Instanzen der gesuchten Relation an die entsprechenden, sie ausdrückenden Stellen annotiert sind. Diese Information wird vom System verwendet, um Regelmuster zu lernen und zu generalisieren. Mithilfe der Information, dass im Satz *„Shakespeare schrieb Hamlet“* die *Autor–von* Relation zwischen *„Shakespeare“* und *„Hamlet“* gilt, kann etwa ein Muster der folgenden Form extrahiert werden: *X schrieb Y*. Ein Beispiel für ein überwachtes, regelbasiertes System ist AUTOSLOG (Riloff, 1996).

Einem halb–überwachten oder Bootstrapping–Ansatz stehen nur einige wenige Ausgangsregeln oder –instanzen zur Verfügung. Da die Ausgangsregeln bzw. –instanzen im englischen als *seed* (englisch: *Saatgut, Saatkorn*) bezeichnet werden, spricht man hierbei auch von Seed–basierten Methoden. Für die *Autor–von*–Relation könnte das Seed aus der Instanz *(„Shakespeare“, „Hamlet“)* bestehen. Damit würde wie im obigen Beispiel das Muster *X schrieb Y* extrahiert werden. Indem man nun alle Instanziierungen X und Y in einem Korpus identifiziert, kann man neue Instanzen der *Autor–von* Relation extrahieren. Diese neuen Instanzen können wiederum dafür verwendet werden, um neue linguistische Muster zu finden, die die betreffende Relation ausdrücken. Dieser Kreislauf wird so lange fortgesetzt, bis ein bestimmtes Konvergenzkriterium erfüllt ist, etwa, dass die neuen Muster einer Iteration keine neuen Instanzen mehr extrahieren. Diese allgemeine Vorgehensweise ist in Abbildung 1 algorithmisch skizziert.

Ein Beispiel eines solchen Ansatzes ist Brin (1999), das ausgehend von einer kleinen Anzahl annotierter Instanzen binärer Relationen Muster extrahiert. Brin (1999) ging es dabei allerdings weniger um die Extraktion von Instanzen aus Fließtext, sondern eher um die Extraktion aus teilweise strukturierten Daten wie HTML–Seiten. Ein ähnliches System, das auf unstrukturierte Texte ausgerichtet ist, ist das SNOWBALL–System (Agichtein und Gravano, 2000).

```
T := Textkorpus
S := Seedinstanzen

WHILE Konvergenzkriterium nicht erreicht DO:
    Finde alle Stellen in T, die eine der Instanzen in S ausdrücken.
    Extrahiere daraus linguistische Muster.
    Verwende die Muster, um neue Instanzen zu finden.
    S = Menge der neu gefundenen Instanzen
END
```

Abbildung 1: Pseudocode für die allgemeine Vorgehensweise bei halb–überwachten Verfahren.

Anpassbarkeit an neue Domänen. Der hohe Aufwand für die manuelle Erstellung von Ressourcen war einer der Gründe für die Verwendung von Machine Learning–Verfahren zur automatischen Erzeugung von Regeln und Mustern. Der Aufwand bezüglich der Anpassung ist zwar abhängig vom Verwandtheitsgrad der beteiligten Relationen, bei manuell erzeugten Regeln ist er jedoch besonders hoch, da jedes Mal, wenn sich die Regeln nicht übernehmen lassen, neue Regeln geschrieben werden müssen. Nur bei eng verwandten neuen Domänen, bei denen der Großteil der bestehenden Regeln gute Ergebnisse liefert, und daher nur wenige neue Regeln hinzugefügt werden müssen, ist die Anpassung relativ einfach. In jedem Fall ist jedoch zumindest ein eingehendes Studium der Extraktionskorpora und oft auch ein gewisses Domänenwissen erforderlich. Darüber hinaus ist es aber auch möglich, dass der Regelformalismus der bestehenden Domäne nicht alle Muster repräsentieren kann, mit denen Instanzen in der neuen Domäne ausgedrückt werden.

Machine Learning–Verfahren reduzieren den notwendigen Aufwand und sind weniger vom genauen Wissen über die linguistischen Strukturen der Zieldomäne abhängig. Bei überwachten Verfahren ist allerdings für eine neue Domäne eine große Zahl annotierter Beispiele nötig, was ebenfalls sehr aufwändig sein kann. Halb–überwachte Verfahren schaffen hier Abhilfe, da sie im Idealfall nur eine angepasste Relationsdefinition und einige wenige Beispielinstanzen erfordern.

2.1.2 Weitere Ansätze

Sequence Labelling. In den letzten Jahren wurde eine Reihe Statistik– und Machine Learning–basierter Verfahren entwickelt und auf das Problem der Relationsextraktion angewendet. Mooney und Bunescu (2005) gruppieren einige dieser Techniken als „sequence–labelling“ Ansätze und unterscheiden dabei statistische Sequenzmodelle von merkmalsbasierten Klassifizierern. Ein Sequence Labelling–Problem ist ein solches, in dem jedem Token

eine Markierung (englisch: *label*) aus einer festen Menge zugewiesen wird. Angewandt auf das Problem der Relationsextraktion bedeutet dies, alle Tokens dahingehend einzuteilen, ob sie Teil einer der Argumentstellen der gesuchten Relation sind.

Statistische Sequenzmodelle stellen eine überwachte Methode dar, die auf einem Korpus annotierter Daten trainiert werden und versuchen, das wahrscheinlichste Labelling einer ganzen Sequenz von Tokens zu finden, etwa eines Satzes. Ein statistisches Sequenzmodell, das erfolgreich bei verschiedenen sprachtechnologischen Problemen Anwendung findet, ist das Hidden Markov Model (HMM). HMMs sind probabilistische endliche Automaten, die besonders gute Ergebnisse liefern, wenn die verwendeten Merkmale größtenteils unabhängig voneinander sind.

Merkmalsbasierte Klassifizierer markieren nicht ganze Sequenzen, sondern klassifizieren jedes Token für sich. Die Merkmale, die für diese Klassifikation verwendet werden, sind jedoch nicht auf dieses Token beschränkt, sondern können Informationen aus dem umgebenden Kontext miteinbeziehen. Hierfür können allgemeine Klassifikationsverfahren angewendet werden. Aus der Machine Learning–Forschung ist eine große Zahl solcher Verfahren bekannt und viele wurden auf Relationsextraktion angewandt. (Für Literaturhinweise zu Beispielen vieler Verfahren siehe: Mooney und Bunescu, 2005).

Klassifikation potentieller Instanzen. Relationsextraktion kann auch als ein binäres Klassifikationsproblem auf höherer Ebene betrachtet werden. Mithilfe eines annotierten Korpus trainieren Zelenko u. a. (2003) einen Klassifizierer, der entscheidet, ob Entitätstupel Instanzen einer bestimmten binären Relation sind oder nicht. In einem Vorverarbeitungsschritt identifiziert ein Named Entity–Tagger alle Erwähnungen von Entitäten im Eingangstext. Hieraus werden dann alle möglichen Instanzen (also Entitätspaare) gebildet, für welche der Klassifizierer daraufhin entscheidet, ob sie tatsächlich Instanzen der Relation sind.

McDonald u. a. (2005) erweitern diesen Ansatz auf Relationen höherer Stelligkeit. Da die Zahl der möglichen Instanzen mit der Zunahme der Stelligkeit exponentiell wächst, zählen sie jedoch nicht alle Tupel auf, sondern trainieren einen Klassifizierer, der, eine bestimmte Relation zugrunde liegend, entscheidet, ob zwei Entitäten bezüglich der Relation miteinander in Beziehung stehen. Aus diesen Beziehungen wird ein Graph aufgebaut, aus dem die Instanzen der gesuchten Relation rekonstruiert werden. Als Instanzen gelten diejenigen Entitätenmengen, innerhalb derer jede Entität mit jeder anderen in Beziehung steht.

Anpassbarkeit an neue Domänen. Die meisten dieser Verfahren können theoretisch ohne Schwierigkeiten auf völlig unbekannte Domänen über-

tragen werden, abgesehen von der eventuell nötigen Anpassung der linguistischen Ressourcen und der Named Entity–Erkennung, die ja auch bei regelbasierten Systemen durchgeführt werden muss. In der Praxis hängen die guten Erkennungsraten jedoch oft von der Feineinstellung einer Vielzahl von Parametern ab, deren Aufwand im Vorfeld schwer abzuschätzen ist. Als ein weiteres Problem könnte sich erweisen, dass bei solchen Methoden die Fehleranalyse, die bei der Anpassung an eine neue Domäne tiefe Einsichten liefern kann, schwierig ist. Erkennungsfehler ergeben sich oft aus dem komplexen Zusammenspiel vieler Parameter oder Merkmale, während bei regelbasierten Systemen oft die für einen Fehler verantwortliche Regel identifiziert werden kann.

2.2 Verarbeitungstiefe

Während die ersten Informationsextraktionssysteme vorwiegend versuchten, ein möglichst vollständiges Textverstehen zu erreichen, woraus die gesuchten Informationen inferiert werden sollten, stützen sich aktuelle Ansätze hauptsächlich auf flache Verarbeitungsmethoden. Heutzutage ist eine große Zahl flacher Analysekomponenten mit hoher Robustheit und Verarbeitungseffizienz auch bei großen Textmengen im Internet frei verfügbar, während dies für tiefere Analysen wie etwa Prädikat–Argumentstrukturzuweisung nur eingeschränkt der Fall ist.

Allerdings wurden in letzter Zeit auch einige erfolgreiche Versuche unternommen, tiefe Verfahren in den Extraktionsprozess einzubinden. Bunescu und Mooney (2007) zeigen den Einfluss, den die Menge der verfügbaren linguistischen Information haben kann, indem sie zwei Systeme vergleichen, die sie entwickelt haben. Eines der beiden berücksichtigt allein die Abfolge der Tokens und deren Eigenschaften, wie etwa die Wortart oder die Position in der WordNet–Hierarchie, während das andere seine Entscheidungen auf die Position der Entitäten im Dependenzgraphen des Satzes stützt. Wie Zhao und Grishman (2005) jedoch bemerken, hat die Entscheidung zwischen flacher und tiefer Verarbeitung in beiden Fällen Nachteile. Flache Methoden bieten nicht so eingehende Informationen über den Text, während tiefe Methoden üblicherweise Schwierigkeit mit der Robustheit haben.

2.3 Modularität

Unter dem Aspekt der Anpassung eines Relationsextraktionssystems auf eine neue Domäne hat die Modularität eines Systems eine große Bedeutung. Unterschiedliche Systeme weisen eine unterschiedliche Verzahnung der einzelnen Komponenten auf. In einigen Ansätzen können sich Teilergebnisse aus einzelnen Komponenten gegenseitig beeinflussen, zum Beispiel Informationen aus der Named Entity–Erkennung und der Relationsextraktion (Roth und Yih, 2002). Das von Miller u. a. (2000) beschriebene System vollführt

Named Entity–Erkennung, syntaktisches Parsing und Relationsextraktion in einem einzigen integrierten Algorithmus. Die Übertragung eines solchen Systems auf eine neue Domäne kann unter Umständen Änderungen am gesamten Algorithmus erfordern und damit auch an Teilen, die sich vielleicht gar nicht ändern müssten, wie etwa dem Parsing. Die einzelnen Komponenten sind hierbei komplexen Wechselwirkungen ausgesetzt, wodurch kleine Änderungen an ihnen ungewollt starken Einfluss auf die Leistung der übrigen Komponenten haben können.

Zhao und Grishman (2005) stellen ein interessantes Beispiel eines integrierten Ansatzes vor, das möglicherweise in geringerem Maße mit diesen Problemen zu kämpfen hat. Sie verwenden Kernelmethoden, um Information aus Tokenisierung, flachem Parsing und Dependenzanalysen zu vereinen. Der Begriff *Kernelmethoden* beschreibt die Generalisierung einer Klasse von Klassifikationsalgorithmen, die normalerweise mit dem Skalarprodukt der Merkmalsvektoren zweier Tokens arbeiten. Sie beinhalten Klassifikationsalgorithmen wie Support Vector Machines oder das Voted Perceptron.

Bei Kernelmethoden wird das Skalarprodukt mit einer speziellen Ähnlichkeitsfunktion ersetzt – einem Kernel. Kernelfunktionen haben die praktische mathematische Eigenschaft, das sie linear kombinierbar sind, wodurch Informationen aus unterschiedlichen Ebenen der linguistischen Verarbeitung in das Klassifikationsergebnis einfließen können, wenn sie als Kernelfunktionen dargestellt werden. Durch diese Kapselung können sie auch leicht ausgetauscht oder verändert werden. Eine Einführung in Kernelmethoden in der Sprachverarbeitung bietet Collins und Duffy (2001).

3 Problemstellung und Motivation

Der Definition von McDonald u. a. (2005) folgend, sei eine *n–stellige Relation* ein Tupel von n Entitätenklassen, wie z. B. *(Person, Datum, Ort)*, wenn man etwa Informationen über den Vortragenden, den Zeitpunkt und den Ort von Präsentationen aus einem Textkorpus extrahieren möchte. Diese Relation stellt zugleich ein *Ereignis* dar. Eine *Instanz* der Relation ist ein Tupel von Entitäten aus diesen Klassen. Für den Satz *„Chomksys Vortrag mit dem Titel 'The Morphophonemics of Modern Hebrew' wird nächsten Donnerstag im Seminarraum stattfinden.“* wäre eine Instanz der Beispielrelation etwa *(„Chomsky“, „nächsten Donnerstag“, „Seminarraum“)*. Arbeitet man mit Relationen höherer Stelligkeiten, ergibt sich häufig die Notwendigkeit, unvollständige Relationsinstanzen darzustellen, da nicht immer alle beteiligten Entitäten im Text erwähnt werden, wie im Satz *„Chomksys Vortrag wird im Seminarraum stattfinden.“*. Fehlende Entitäten werden durch $\perp$ repräsentiert, für das eben erwähnte Beispiel erhielte man also *(„Chomsky“,* $\perp$*, „Seminarraum“)*.

Die Domäne der Relationsextraktion ist zum Teil durch die gewählte Re-

lationsdefinition vorgegeben, da sie die Menge der extrahierbaren Instanzen bereits einschränkt. Dies allein reicht jedoch nicht aus, da unsere Beispielrelation anstelle von Präsentationsterminen auch Geburtstage beschreiben könnte. Die gesuchten Instanzen müssen also noch enger vorgegeben werden. Bei überwachten Methoden sind diese in den Trainingsdaten annotiert, bei unüberwachten sind sie durch die Seed–Instanzen gegeben. Diese Instanzen geben implizit zu einem gewissen Grad die Arten von Text vor, in denen sie vorkommen. Da Relationen jedoch oft in ganz unterschiedlichen Texten vorkommen können – für Geburtstage sind etwa Nachrichtentexte, persönliche Homepages oder Lebensläufe denkbar – leistet die Beschaffenheit des Extraktionskorpus ebenfalls einen Beitrag zur Definition der Domäne.

Bei der Einschätzung des Aufwands zur Anpassung an eine neue Domäne muss unterschieden werden zwischen dem Aufwand, der für die Definition und Verarbeitung der neuen Domäne anfällt und den Änderungen, die am eigentlichen System durchgeführt werden müssen. Für eine neue Domäne sind in jedem Falle eine Relationsdefinition, ein Korpus und die Definition der erwarteten Ergebnisse für eine Evaluation (siehe hierzu auch Abschnitt 5) vonnöten. Dieser Aufwand ist unabhängig von den Änderungen am Extraktionssystem, die das Schreiben neuer Regeln, das Anpassen von Parametern und Algorithmen oder das eventuell nötige Anpassen (bei integrierten Systemen) oder Austauschen (bei modularen Systemen) von linguistischen Komponenten wie Tokenisierung, Named Entity–Erkennung oder Parsing beinhalten können.

Keines der in Abschnitt 2 vorgestellten Systeme lässt sich ganz ohne Anpassungen auf eine neue Domäne übertragen, der nötige Aufwand an Anpassungen ist jedoch zum Teil sehr unterschiedlich. Manuell erstellte Regeln bedeuten einen sehr hohen Aufwand, vergleichbar mit dem Aufwand, der für die Erstellung des Systems für die Ursprungsdomäne betrieben werde musste. Für ein überwachtes System, das Regeln automatisch lernt, sind für die Zieldomäne annotierte Daten nötig. Da nicht–regelbasierte Systeme nur in eingeschränkterem Maße Einsichten über die Unterschiede zwischen Domänen zulassen als es bei regelbasierten der Fall ist, stellt das System von Hong (2006) die Grundlage für die vorliegenden Experimente dar.

Dieses System arbeitet seed–basiert, ist also halb–überwacht, und extrahiert Regeln und Relations– bzw. Ereignisinstanzen in einem iterativen Algorithmus, der in Abschnitt 4 ausführlich erläutert wird. Es weist einige Eigenschaften auf, die es für die Anpassung an neue Domänen gut eignen. Zum einen verwendetet es Relationsinstanzen als Seed und nicht Regeln, es ist also semantisch orientiert und nicht syntaktisch, da das Seed keine Information über die erwartete Struktur des Textes enthält, sondern allein Informationen über die beteiligten Entitäten. Dies ist für die Übertragung auf neue Domänen hilfreich, da es bei Syntax–orientierten Ansätzen vorkommen kann, dass die syntaktischen Muster der Zieldomäne nicht mit dem Repräsentationsschema der Ursprungsdomäne ausgedrückt werden können.

Bei der Verwendung von Instanzen ist dies nicht der Fall. Zusätzlich dazu ist das System modular aufgebaut. Es verwendet Named Entity–Erkennungs– und Parsingkomponenten, die mit geringem Aufwand ausgetauscht werden können, wenn die Zieldomäne dies erfordert.

Die vielversprechenden Ergebnisse dieses Systems bei der Extraktion von Preisverleihungsereignissen in der Nobelpreisdomäne (Xu u. a., 2007; Hong, 2006) haben die Anwendung des Systems auf die verwandte Domäne der Musikpreise motiviert. In der vorliegenden Arbeit werden Instanzen der Relation *(Künstler, Jahr, Preis, Preisgebiet)* in dieser Domäne gesucht. Dies stellt aus zweierlei Gründen ein interessantes Problem dar. Zum einen ist die Domäne der Musikpreise interessant, weil sie sich von der Nobelpreisdomäne insofern unterscheidet, dass Erwähnungen von Instanzen seltener sind (siehe Abschnitt 5.2), und dass die Redundanz der Erwähnungen sehr gering ist. Das Verhältnis von Erwähnungen zu erwähnten Instanzen ist sehr gering. Auf der anderen Seite sind die beiden Probleme so sehr miteinander verwandt – es geht in beiden Fällen um Preisverleihungen – dass es nahe liegt, die Übertragung der Extraktionsregeln aus der Nobelpreisdomäne zu versuchen und zu evaluieren. Die Regeln, die für die weitaus prominenteren Nobelpreise gefunden wurden, können für die Musikpreisdomäne von Nutzen sein.

4 Lösungsansatz

Der hier vorgestellte Ansatz basiert auf Xu u. a. (2006). Deren Ansatz arbeitet halb–überwacht mit Semantik–orientierten Seeds auf einer unannotierten Menge freien Textes. Der Extraktions– und Lernalgorithmus kann auf folgende Weise skizziert werden:

1. Auffinden der Seedinstanzen im Extraktionskorpus und Annotation der Seedargumenterwähnungen im Text.
2. Lernen von Mustern und Regeln aus den Textsegmenten, die die Seedargumente enthalten.
3. Anwenden der gewonnenen Regeln auf das Textkorpus, um neue Instanzen zu erhalten.
4. Wiederholen des Algorithmus ab 1., mit den neu extrahierten Instanzen als Seed.
   Abbruch, sobald die neu gelernten Regeln keine neuen Instanzen mehr extrahieren.

4.1 Implementierung

Die vorliegende Arbeit beschreibt eine Erweiterung des Systems von Hong (2006), welches das Verfahren von Xu u. a. (2006) implementiert. Auf dieser

Grundlage wurde ein System entwickelt, welches die Extraktion von Musikpreisverleihungen realisiert. Im Folgenden wird zunächst die Arbeitsweise des Systems von Hong (2006) beschrieben, welches für die Nobelpreisdomäne entwickelt wurde, und anschließend die vorgenommenen Erweiterungen und Anpassungen erläutert.

4.1.1 Verwendete Werkzeuge

In der Implementierung des Systems wurden einige existierende Textverarbeitungswerkzeuge verwendet, die hier kurz beschrieben werden sollen.

Lucene Lucene[2] ist eine Open Source–Bibliothek zur Indizierung von und Suche in Dokumentensammlungen. Ein Lucene–Index besteht aus einer Menge von 'Dokumenten', die ihrerseits eine Menge von 'Feldern' enthalten, in denen bestimmte Daten gespeichert sind. Die Art der Dokumente und die verwendeten Felder werden weiter unten beschrieben.

JToken JToken ist ein Tokenisierungswerkzeug, mit dessen Hilfe Fließtext in einzelne Teile zerlegt werden kann. Diese Teile können Absätze, Sätze, Worttokens oder andere Einheiten sein. In diesem System wird JToken verwendet, um Dokumente in Sätze aufzugliedern.

SProUT SProUT – *Shallow Processing with Unification and Typed feature structures* (Drozdzynski u. a., 2004) ist eine Entwicklungsplatform für multilinguale Textverarbeitungssysteme, die getypte Merkmalsstrukturen und Unifikations–basierte Regeln unterstützt. In diesem System wird die SProUT–eigene Named Entity–Erkennungskomponente verwendet und erweitert, um Erwähnungen von Preisgewinnern, Preisen, Preisgebieten und Jahren zu erkennen und zu markieren.

Minipar Minipar (Lin, 1998) ist ein effizienter und robuster Dependenzparser. Die von Minipar erzeugten Dependenzgraphen bilden die Grundlage für das Lernen syntaktischer Muster und Regeln und die Regelanwendung.

4.1.2 Vorverarbeitung

Vor der eigentlichen Anwendung des Systems sind einige Vorarbeiten nötig, die die benötigten Ausgangsdaten aufbereiten.

Korpus und Seed Um die Extraktion durchzuführen, muss eine Sammlung von Fließtext vorhanden sein, die Erwähnungen der gesuchten Relation bzw. des gesuchten Ereignisses enthalten. Desweiteren sind einige wenige

[2] `lucene.apache.org`

Seedinstanzen der Relation erforderlich. Eine Beschreibung dieser Daten und wie sie gesammelt wurden findet sich in den Abschnitten 5.1 und 5.2.

Indizierung der Dokumente und Sätze Die Dokumente im Extraktionskorpus werden zur besseren Durchsuchbarkeit und zur einfacheren Handhabung mit LUCENE indiziert. Der Text eines Dokuments ist dabei über das Feld *text* zugänglich, außerdem enthält der Index die Felder *url* (URL des Artikels), *year* (Erscheinungsjahr des Artikels), *title* (Titel des Artikels) und einige weitere.

Anschließend wird jedes Dokument mithilfe von JTOKEN in Sätze zerlegt und die Sätze über einen Satzindex zugänglich gemacht. Hierbei enthält das Feld *text* den Satz selbst, *year* das Erscheinungsjahr des Artikels, der den Satz enthält, *url* die URL und *path* eine eindeutige Bezeichnung der Quelle des Artikels aus dem der Satz stammt.

4.1.3 Implementierung des Algorithmus

Beschrieben wird im Folgenden eine Iteration des Algorithmus. Dies wird solange wiederholt, bis in einer Iteration keine neue Instanzen mehr extrahiert werden, wobei das „Seed“ ab der $n + 1$–ten Iteration aus den in der n–ten Iteration gewonnenen neuen Instanzen besteht.

Zuerst werden für jede Seedinstanz relevante Dokumente gesucht. Ein Dokument gilt als relevant, wenn es mindestens Erwähnungen des Preisträgers, des Preises und des Preisgebietes enthält. Alle Sätze in diesen Dokumenten werden nun daraufhin überprüft, ob sie für die Instanz relevant sind. Hierfür werden zunächst alle Entitätserwähnungen in diesen Sätzen mit SPROUT markiert und diese Erwähnungen mit den Argumenten der Seedinstanz verglichen. Gibt es dabei ausreichende Übereinstimmungen, so gilt der Satz als relevant. Für die Nobelpreisdomäne ist dies der Fall, wenn der Satz mindestens drei Argumente der Instanz enthält.

Für die so gewonnenen Sätze werden mit MINIPAR Dependenzgraphen konstruiert, an deren Knoten die Ausdehnung der jeweiligen Entitätserwähnungen annotiert wird. Aus diesen Graphen werden nun als relevante syntaktische Muster die kleinstmöglichen Teilgraphen extrahiert, die die Argumente der Instanz enthalten. Abbildung 2 illustriert diesen Vorgang. Die Wurzel eines solchen Musters unterliegt dabei der Einschränkung, dass sie ein lexikalischer Knoten sein muss (also ein Wort im Satz bezeichnet, im Gegensatz zu funktionalen Knoten wie *fin*, die von MINIPAR erzeugt werden). Aus den extrahierten Mustern werden Regeln generiert, die zum Auffinden neuer Instanzen verwendet werden. Bei der Überführung von Mustern in Regeln werden ähnliche Muster miteinander kombiniert. Abbildung 3 zeigt ein Beispiel für eine Regel. Eine Regel besteht aus einem Kopf und eventuell einigen Unterregeln. Der Kopf kann entweder ein Wortstamm oder ein Platzhalter für ein Argument der Relation sein. Die Unterregeln sind wiederum

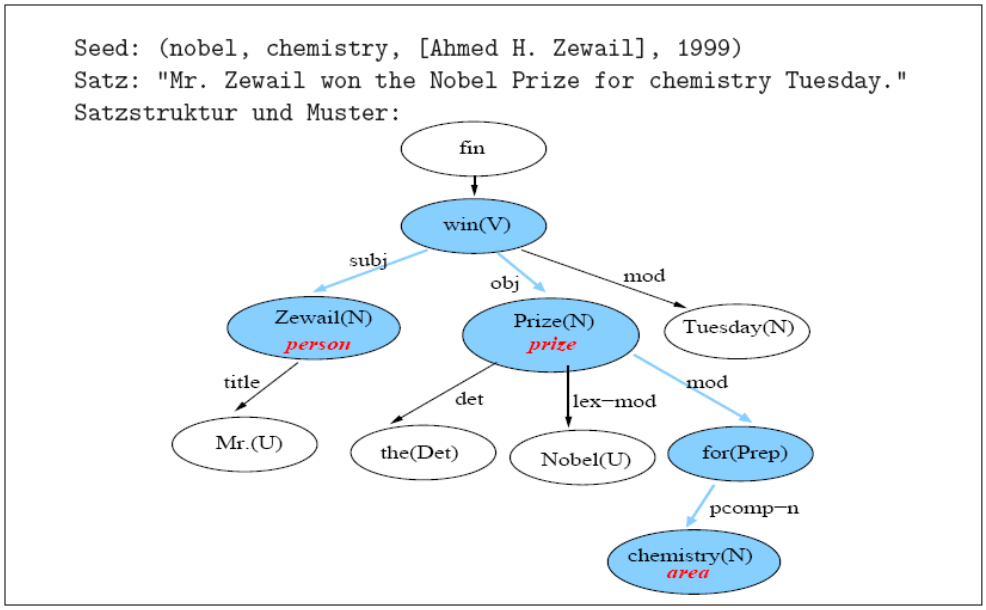

Abbildung 2: Beispiel für die Extraktion eines Musters aus einem Dependenzgraphen für einen Satz. Die farblich markierten Knoten entsprechen dem Muster. *Quelle: Hong (2006)*

Regeln, die aber mit der sie enthaltenden Regel über Relationen verbunden sind. Diese Relationen entsprechen den Dependenzbeziehungen, die bei der Regelanwendung zwischen den Knoten im Dependenzgraphen bestehen müssen, auf die die Köpfe der Regel und ihrer Unterregeln passen, damit die gesamte Regel passt. Eine detailliertere Beschreibung dieser Vorgänge findet sich in Hong (2006).

In einem zweiten Teil des Algorithmus werden die neuen Regeln auf das Korpus angewendet. Um nicht alle Sätze untersuchen zu müssen, werden sie anhand einer Suchanfrage an den Satzindex grob vorausgewählt. Zunächst

```
{ head(("win", V))
    -> subj( { head([winners]) } )
    -> obj( { head([prize])
                -> mod( { head(("in", Prep))
                            -> pcomp-n( { head([area]) } )
                        } )
            } )
}
```

Abbildung 3: Beispiel einer Extraktionsregel

müssen die Sätze mit Dependenzgraphen und Entitätserwähnungen annotiert werden. Dann wird für jeden Satz jede Regel angewendet. Die Anwendung einer Regel ist erfolgreich, genau dann, wenn ein Knoten existiert, auf den der Regelkopf passt, und dessen Unterregeln auf je einen seiner Kindknoten im Dependenzgraphen erfolgreich angewendet werden können, und diese Kindknoten mit dem Knoten in der in der Regel definierten Beziehung stehen. Passen mehrere Regeln auf einen Satz, werden die dabei extrahierten Instanzen, wenn möglich, vereinigt.

Die bei der Anwendung der Regeln gefundenen Instanzen werden wie bereits beschrieben in der nächsten Iteration als Seed verwendet.

4.2 Anpassungen und Erweiterungen

SProUT Ein Teil der nötigen Erweiterungen fand an der SProUT–Grammatik statt. Die Erkennung der Erwähnungen von Jahresdaten erfolgt durch eine in SProUT enthaltene Grammatik. Für Künstler, Preise und Preisgebiete mussten neue Teilgrammatiken entwickelt werden. Dies geschah mittels sogenannter Gazetteer–Listen, die im Grunde einfach Listen von Oberflächenformen darstellen. Beispielhafte Gazetteer–Einträge sehen folgendermaßen aus:

```
eminem | GTYPE:gaz_music_artist | CONCEPT:eminem | LANG:any
marshall mathers | GTYPE:gaz_music_artist | CONCEPT:eminem | LANG:any

grammy award | GTYPE:gaz_music_award | CONCEPT:grammy_award | LANG:en
mtv vma | GTYPE:gaz_music_award | CONCEPT:mtv_video_music_award | LANG:en
razzie | GTYPE:gaz_music_award | CONCEPT:golden_raspberry_award | LANG:en
```

Jede Zeile beschreibt eine Merkmalsstruktur, wobei die einzelnen Merkmale durch `|` getrennt sind. Die an erster Stelle stehende Zeichenkette bezeichnet diejenige Oberflächenform, der die entsprechende Merkmalsstruktur zugewiesen werden soll. Das Attribut `GTYPE` steht für *gazetteer type*, also den allgemeinen Typen des Eintrages. Die verwendeten Typen sind `gaz_music_artist`, `gaz_music_prize` und `gaz_music_prize_area`. In unserem Fall sind Gazetteertypen und Named Entity–Typen zueinander isomorph. SProUT erlaubt aber auch komplexere Definitionen von Named Entity–Typen. Das Merkmal `CONCEPT` beschreibt den semantischen Inhalt, der der Zeichenkette zugewiesen wird, und der in der Darstellung und Verarbeitung von Instanzen verwendet wird, um von Oberflächenformen zu abstrahieren.

Diese Einträge wurden von Hand erstellt, wobei die Oberflächenformen durch Introspektion, Korpusrecherche und durch die Verwendung der unterschiedlichen Formen, die in den Beschreibungen gleicher Instanzen in unter-

schiedlichen Quellen bei der Recherche des Ideal Table (siehe Abschnitt 5.1) vorkamen, gewonnen wurden.

Variantenindex Es fällt auf, dass eine Umkehrung eines Teils der Gazetteer–Listen, genauer gesagt eine Abbildung von Konzepten auf Mengen von Oberflächenformen, einem Konzept alle möglichen, in der Gazetteer–Liste berücksichtigten, Oberflächenvarianten zuordnet. Dies ist bei der Häufigkeit der Verwendung von Namensvarianten in der Musikdomäne, sowohl bei Künstlern, als auch bei Preisen und Gebieten, von großer Bedeutung. Der Variantenindex findet vor allem bei der Suche nach relevanten Dokumenten für Seedinstanzen Anwendung. Hierfür enthält er zu jedem Konzept bereits eine Lucene–Suchanfrage, die aus einer Disjunktion der Varianten besteht.

Auswahl relevanter Sätze Für die Musikpreisdomäne gilt: Enthält ein Satz mindestens zwei Argumente der Instanz, wovon eines der Preisgewinner ist, gilt er für die Musikpreisdomäne als relevant. Der Grund hierfür erschließt sich aus der Beschaffenheit des Textkorpus. Siehe hierzu auch Abschnitt 5.2, Tabelle 6.

Auswahl von Sätzen für die Regelanwendung Bei Hong (2006) wird die Anfrage *„nobel“* verwendet, um Sätze für die Regelanwendung vorauszuwählen. Da es für die Musikpreisdomäne kein ähnlich prominentes geeignetes Wort gibt, werden mithilfe des Variantenindexes Sätze gesucht, die entweder einen Preis oder einen Künstler erwähnen. Dank der Indizierung der Sätze mit Lucene geht dies sehr effizient.

Regeln als Seed Um die Regeln aus der Nobelpreisdomäne übertragen zu können, wurde dem System die Möglichkeit hinzugefügt, auf der Basis eines Regelseeds anstelle eines Instanzseeds zu beginnen. Dabei werden die übertragenen Regeln zunächst einmal angewendet. Das Ergebnis dieser Anwendung dient dem System dann als Instanzseed für die erste Iteration.

5 Ergebnisse

Für die vorliegende Arbeit wurden zwei Extraktionsexperimente durchgeführt und verglichen. Im ersten Experiment wurde das System auf der Basis eines Instanzseeds angewendet, im zweiten auf der Basis der übertragenen Regeln. Die folgenden Abschnitte erläutern die Ergebnisse dieser Experimente und ihre Auswertung.

5.1 Goldstandard

Bei der Evaluation halb–überwachter Methoden stellt sich generell das Problem der Definition eines Goldstandards. Da der Wunsch nach dem Verzicht auf annotierte Korpora häufig ein Anlass für die Verwendung solcher Methoden ist, stehen im Allgemeinen keine Annotationen zur Verfügung, aus denen sich das perfekte hypothetische Ergebnis einer Extraktion ableiten lässt, welches normalerweise zum Vergleich mit den tatsächlichen Ergebnissen und der Bildung der Evaluationsmaße *Precision* und *Recall* herangezogen wird. Dies ist für diese Arbeit ebenfalls zutreffend.

Agichtein und Gravano (2000) folgend, kann man sich Abhilfe schaffen mit der Erstellung eines sogenannten Ideal Table, der alle tatsächlichen Instanzen des gesuchten Ereignisses auflistet. Dadurch lässt sich die Korrektheit (*Precision*) der Ergebnisse errechnen und man erhält eine obere Grenze für die Zahl der extrahierbaren korrekten Instanzen, auch wenn unklar bleibt, wie viele davon tatsächlich im untersuchten Text enthalten sind. Dies ist natürlich nur gangbar, wenn die Zahl der Instanzen endlich und im Voraus abschätzbar ist. Im Falle von Xu u. a. (2006), wo das Ziel der Extraktion Instanzen des Ereignisses *Nobelpreisverleihung* ist, existieren strukturierte Datenbanken und Listen mit diesen Instanzen. Um ein ähnliches Instrument für *Musikpreise* zu erhalten, ist ein anderes Vorgehen nötig. Hierbei geht es um eine ganze Reihe von Preisen und möglichen Gewinnern, und die Anzahl der Instanzen ist nicht leicht vorhersagbar. Daher habe ich mich in der vorliegenden Arbeit auf die in Tabelle 1 angegebenen populären Künstler und Preise und die dadurch vorgegebenen Preiskategorien beschränkt.

Künstler:	*Preise*:	
Eminem	Academy Awards (Oscars)	Grammy Awards
Madonna	American Music Awards	Meteor Ireland Music Awards
U2	Billboard Awards	MTV Video Music Awards
	Brit Awards	MTV European Music Awards
	Golden Globe Awards	MTV Movie Awards
	Golden Raspberry Awards	The Source Awards

Tabelle 1: Untersuchte Musiker und Preise.

Für einen Teil dieser Preise, wie z.B. die Grammy Awards, stehen offizielle Siegerlisten zur Verfügung, für andere mussten die Instanzen aus verschiedenen Sekundärquellen[3] zusammengetragen werden, deren Korrektheit nicht so verlässlich ist wie die offizieller Quellen. Allerdings wurden die Instanzen stets aus mehreren verschiedenen Quellen verglichen und manuell zusammengeführt, was eine gewisse Verlässlichkeit gewährleistet. Insgesamt wurden 168

[3]Besonders hilfreich waren hierbei `en.wikipedia.org` und `www.rockonthenet.com`. Weitere Quellen waren u.a. `www.absolutemadonna.com`, `www.aceshowbiz.com`, `www.eminem.net`, `www.imdb.com`, `www.trshady.com` und `www.u2faqs.com`.

Instanzen von Preisverleihungen in den Ideal Table aufgenommen, die Verteilung auf Künstler, Preise und Jahre ist in Tabelle 2 aufgeführt. Eine Liste aller Instanzen findet sich in Anhang A.

Künstler:	
Eminem	55
Madonna	56
U2	57

Preise:	
Academy Awards (Oscars)	3
American Music Awards	11
Billboard Awards	12
Brit Awards	14
Golden Globe Awards	7
Golden Raspberry Awards	14
Grammy Awards	36
Meteor Ireland Music Awards	17
MTV Video Music Awards	32
MTV European Music Awards	15
MTV Movie Awards	4
The Source Awards	3

Jahre:			
1983	1	1996	2
1986	1	1997	6
1987	4	1998	9
1988	4	1999	6
1989	8	2000	17
1990	4	2001	21
1991	2	2003	26
1992	4	2004	4
1993	4	2005	5
1994	2	2006	10
1995	4	2007	1

Tabelle 2: Verteilung der Künstler, Preise und Jahre im Ideal Table.

5.2 Extraktionskorpus

Das für die Extraktion verwendete Korpus enthält Nachrichtentext, der sich mit Verleihungen von Musikpreisen befasst. Insgesamt besteht es aus 324.479 Sätzen in 8.920 Dokumenten (Artikeln) der BBC Webpräsenz[4], die über die BBC–eigene Suchmaske[5] gefunden wurden. Die hierfür verwendeten Suchanfragen waren verschiedene Kombinationen von Künstler–, Preis– und Preisgebietsnamen, wobei erneut der in Abschnitt 4 vorgestellte Variantenindex verwendet wurde.

n	*Dokumente*	*potentiell relevant*	%
2	3896	2492	$63,96\%$
3	1260	414	$32,86\%$
4	221	63	$28,51\%$

Tabelle 3: Verteilung der Dokumente mit je n Entitätstypen.

Für die Regelanwendung wurden die 14.076 Sätze verwendet, die mindestens eine Erwähnung eines Preises oder eines Künstlers enthalten. Tabellen 3 und 4 zeigen die Anzahl von Dokumenten bzw. Sätzen, die mehrere Entitätserwähnungen enthalten, zusammen mit dem Anteil der jeweils potentiell relevanten Dokumente oder Sätze. Als potentiell relevant gelten sie, wenn sie

[4] `news.bbc.co.uk`
[5] `search.bbc.co.uk`

n	*Sätze*	*potentiell relevant*	%
2	1993	933	$46,83\%$
3	152	43	$28,29\%$
4	6	4	$66,66\%$

Tabelle 4: Verteilung der Sätze mit je n Entitätstypen.

eine Entitätskombination enthalten, wie sie in einer Instanz im Ideal Table vorkommt. Sie sind nur *potentiell* relevant, da allein das Vorkommen der korrekten Entitäten nicht hinreichend für eine tatsächliche Instanz ist. Häufig kommt es vor, dass über Nominierungen berichtet wird, oder dass neben den Gewinnern eines Preises auch diejenigen Künstler erwähnt werden, gegen die sie sich durchgesetzt haben. In den Tabellen 5 und 6 sind die Dokumente bzw. Sätze weiter aufgeschlüsselt nach den Typen der Entitäten, die sie enthalten.

Entitätstypen	*Dokumente*	*potentiell relevant*	%
Künstler, Preis, Gebiet, Jahr	221	63	$28,51\%$
Künstler, Preis, Gebiet	329	195	$59,27\%$
Künstler, Preis, Jahr	481	206	$42,83\%$
Künstler, Gebiet, Jahr	315	127	$40,32\%$
Preis, Gebiet, Jahr	798	141	$17,67\%$
Künstler, Preis	750	672	$89,60\%$
Künstler, Gebiet	516	434	$84,11\%$
Künstler, Jahr	1176	1026	$87,24\%$
Preis, Gebiet	1256	517	$41,16\%$
Preis, Jahr	2324	1156	$49,74\%$
Gebiet, Jahr	1057	411	$38,88\%$

Tabelle 5: Verteilung der Kombinationen von Entitätstypen in Dokumenten.

Für diese vier Tabellen gilt, dass ein Dokument oder Satz, das bzw. der eine bestimmte Anzahl an Entitäten enthält, für die jeweils kleineren Anzahlen erneut gezählt wurde. Ein Dokument, dass etwa für vier Entitäten einmal gezählt wurde, wurde für drei Entitäten viermal ($\hat{=} \binom{4}{3}$ Kombinationen) gezählt und für zwei Entitäten sechsmal ($\hat{=} \binom{4}{2}$ Kombinationen). Dies erleichtert die spätere Evaluation und erklärt die hohen Zahlen der Vorkommen zweier und dreier Entitäten. Bei der Beurteilung dieser Daten sollte beachtet werden, dass das untersuchte Korpus nur aufgrund von Suchanfragen für drei Künstler erstellt wurde und daher besonders unter diesem Gesichtspunkt kein balanciertes Abbild der Domäne der Musikpreisverleihungen darstellt, sondern lediglich einen eingeschränkten Ausschnitt.

Diese Daten deuten an, dass nur über einen Bruchteil der Instanzen im Ideal Table tatsächlich im Extraktionskorpus berichtet wird. Tabelle 7 zeigt,

Entitätstypen	*Sätze*	*potentiell relevant*	%
Künstler, Preis, Gebiet, Jahr	6	4	$66,66\%$
Künstler, Preis, Gebiet	27	19	$70,37\%$
Künstler, Preis, Jahr	33	20	$60,60\%$
Künstler, Gebiet, Jahr	12	8	$66,66\%$
Preis, Gebiet, Jahr	98	8	$8,16\%$
Künstler, Preis	195	188	$96,41\%$
Künstler, Gebiet	196	176	$89,80\%$
Künstler, Jahr	256	207	$80,86\%$
Preis, Gebiet	617	158	$25,61\%$
Preis, Jahr	849	302	$35,57\%$
Gebiet, Jahr	202	54	$26,73\%$

Tabelle 6: Verteilung der Kombinationen von Entitätstypen in Sätzen.

n	*Instanzen*
2	168
3	151
4	55

Tabelle 7: Zahlen der Instanzen im Ideal Table, für die n ihrer Entitäten innerhalb eines Dokuments gefunden wurden.

für wie viele Instanzen das Korpus Dokumente enthält, in denen die Entitäten dieser Instanz vorkommen, untergliedert nach der Anzahl der Instanzen. Tabelle 8 zeigt dasselbe für Sätze. Für die Evaluation der Extraktionsergebnisse werden nur die Instanzen aus dem Ideal Table verwendet, für die sich Sätze im Korpus befinden, die dieselben Entitätskombinationen enthalten, die sich auch in der Instanz finden. Diese stellen zwar nun keine allgemeine Obergrenze für die aus dem Text extrahierbaren Instanzen mehr dar, da der Named Entity–Erkennung höchstwahrscheinlich einige Erwähnungen entgehen und Dokumente auch ohne wörtliche Erwähnung der Entitäten von den Instanzen berichten könnten. Da der vorgestellte Ansatz dies jedoch ohnehin nicht berücksichtigt, bieten die so erhaltenen Maße ein aussagekräftigeres Bild über die Qualität der Extraktion.

n	*Instanzen*
2	168
3	49
4	5

Tabelle 8: Zahlen der Instanzen im Ideal Table, für die n ihrer Entitäten innerhalb eines Satzes gefunden wurden.

5.3 Ergebnisse der Extraktion

Das modifizierte System wurde wie in Abschnitt 3 und 4 beschrieben auf das Korpus angewendet, einmal unter Verwendung von Seedinstanzen und einmal unter Verwendung von Extraktionsregeln als Seed, die bei der Relationsextraktion in der Nobelpreisdomäne erhalten wurden.

5.3.1 Instanzen als Seed

Stelligkeit	*extrahierte Instanzen*	*davon im Ideal Table*	*Abdeckung des Ideal Table*[6]
2	17	$76,47\%$	$10,12\%$
3	2	$100,00\%$	$4,08\%$
4	0	–	–

Tabelle 9: Ergebnisse bei der Verwendung von Instanzen als Seed.

Entitätstypen	*extrahierte Instanzen*	*davon im Ideal Table*
Künstler, Preis, Jahr	1	$100,00\%$
Künstler, Preis, Gebiet	1	$100,00\%$
Künstler, Jahr	10	$70,00\%$
Künstler, Preis	7	$85,71\%$

Tabelle 10: Ergebnisse bei der Verwendung von Instanzen als Seed nach Typenkombinationen aufgeschlüsselt.

Die Ergebnisse der Extraktion mit Instanzen als Seed sind in Tabellen 9 und 10 angegeben. In dieser und den folgenden Tabellen bezeichnet *Abdeckung des Ideal Table* den Anteil der erwarteten Teilmengen des Ideal Table (siehe 5.2), der in den gefundenen Instanzen enthalten ist. Analog zur Erstellung der Statistiken in 5.2 wurden gefundene höherstufige Instanzen ebenfalls für niederstufige gewertet: vierstellige Instanzen zählen ebenfalls als die in ihnen als Teilmenge enthaltenen dreistelligen und zweistelligen Instanzen, dreistellige ebenfalls als zweistellige. In der Spalte *extrahierte Instanzen* sind allerdings nur die tatsächlich in dieser Stelligkeit extrahierten Instanzen aufgeführt.

Als Seed wurde die Instanz *(Eminem, 2001 Grammy award, Best Rap Album)* gewählt, was experimentell zu den besten Ergebnissen führte. Insgesamt wurden bei diesem Vorgehen 115 Regeln extrahiert, davon 108 2–stellige, also solche, die zwei Relationsargumente extrahieren, 7 3–stellige und keine 4–stellige. Tabelle 11 zeigt die Verteilung auf Typenkombinationen.

[6] Bezieht sich auf die Teilmenge des Ideal Table für die im Korpus Dokumente enthalten sind, siehe 5.2. Hinweise zur Berechnung siehe 5.3.1.

Entitätstypen	*extrahierte Regeln*
Künstler, Preis, Gebiet	4
Künstler, Preis, Jahr	3
Künstler, Preis	66
Künstler, Jahr	34
Künstler, Gebiet	6
Preis, Jahr	4

Tabelle 11: Ergebnisse bei der Verwendung von Instanzen als Seed nach Typenkombinationen aufgeschlüsselt.

5.3.2 Übertragene Extraktionsregeln

Für das Experiment zur Übertragung der Extraktionsgrammatik aus der Nobelpreisdomäne wurden 717 Regeln übertragen, davon 131 4–stellige, 452 3–stellige und 134 2–stellige. Tabelle 12 zeigt die Ergebnisse, die bei der Verwendung von Regeln anstelle von Instanzen als Seed erzielt werden.

Stelligkeit	*extrahierte Instanzen*	*davon im Ideal Table*	*Abdeckung des Ideal Table*[6]
2	36	$75,00\%$	$17,26\%$
3	2	$50,00\%$	$2,04\%$
4	0	–	–

Tabelle 12: Ergebnisse bei der Verwendung von Regeln als Seed.

Entitätstypen	*extrahierte Instanzen*	*davon im Ideal Table*
Künstler, Preis, Gebiet	2	$50,00\%$
Künstler, Jahr	10	$70,00\%$
Künstler, Preis	9	$88,89\%$
Künstler, Gebiet	17	$70,59\%$

Tabelle 13: Ergebnisse bei der Verwendung von Regeln als Seed nach Typenkombinationen aufgeschlüsselt.

Es eignet sich offenbar nur ein Teil der Regeln für die Übertragung. Tabelle 14 zeigt, welche Regeln wie wie oft angewendet werden konnten. Zusätzlich zu den übertragenen Regeln wurden 157 Regeln extrahiert, davon 3 3–stellige und 154 2–stellige. Die übertragenen Regeln sind für ca. zwei Drittel der extrahierten Instanzen verantwortlich.

5.4 Diskussion

Betrachtet man die extrahierten Instanzen, zeigt sich deutlich, welchen Vorteil die Wiederverwendung einer Extraktionsgrammatik haben kann. Im vorliegenden Fall, in dem nur wenige Instanzen extrahiert werden können, und

erfolgreiche Anwendungen	*Regelkopf*	*Entitätstypen*
8	("leader", N)	Künstler, Jahr
4	[Künstler]	Künstler, Preis
4	[Künstler]	Künstler, Gebiet
3	("win", V)	Künstler, Preis
1	("win", V)	Künstler, Preis, Gebiet
1	("be", VBE)	Künstler, Jahr
1	[Künstler]	Künstler, Preis
1	[Künstler]	Künstler, Gebiet
1	[Preis]	Preis, Gebiet
1	[Gebiet]	Künstler, Gebiet
1	[Gebiet]	Künstler, Gebiet

Tabelle 14: Anwendbarkeit übertragener Regeln.

deshalb nur wenige brauchbare Extraktionsregeln generiert werden, hat die Verwendung der Nobelpreisregeln zu einer Verdopplung der extrahierten Instanzen bei gleicher Qualität geführt. Dies gilt zumindest für 2–stellige Instanzen. An 3–stelligen Instanzen wurden allerdings so wenige extrahiert, dass die Unterschiede zu vernachlässigen sind.

Das System zeigt gute Ergebnisse, was die Korrektheit der extrahierten Instanzen anbelangt, findet allerdings nur einen sehr geringen Teil der möglicherweise im Text enthaltenen Instanzen und nur Instanzen geringer Stelligkeit. Eine mögliche Erklärung für die geringe Abdeckung des Ideal Table ist – abgesehen von der Tatsache, dass nur ein Teil der Instanzen im Korpus erwähnt wird – die zu große Spezifizität der meisten Regeln. Zu einem gewissen Anteil ist dies auf Fehler in den von MINIPAR generierten Dependenzgraphen zurückzuführen, wie in den beiden Beispielen in Abbildung 4 und 5. Das zweite Beispiel illustriert noch zwei weitere Punkte, die das Auffinden der Instanzen erschweren. Zum einen werden sehr häufig Koordinationen verwendet, um mehrere Relationen auf einmal auszudrücken. Die Zuordnung von Entitäten zueinander ist hierbei besonders kompliziert. Aber auch die syntaktische Verarbeitung steht oft vor einem Problem und generiert falsche Analysen. Ein weiterer Punkt, der hier nur angedeutet ist, ist die häufige metonyme Verwendung von Werken der Künstler (*„The Marshall Mathers LP“*) anstelle der Künstler selbst. Ein besseres Beispiel hierfür ist der Satz *„The Marshall Mathers LP won a Grammy Award for best rap album in 2000.“*. Um mit solchen Phänomenen umzugehen, ist ein ausgefeilteres Erkennen von Entitäten erforderlich als in der hier vorgestellten Arbeit. Dies schließt auch weitere Maßnahmen zur Erhöhung der Abdeckung, wie eine robuste Koreferenzanalyse, mit ein, denn wie Doddington u. a. (2004) bemerken, geht Named Entity–Erkennung weit über die bloße Erkennung von Namen hinaus.

Ein weiteres Problem ist, dass das System in seiner akutellen Implemen-

```
{ head([prize])
    -> nn( { head(("Eminem won two", N)) } )
                -> lex-mod( { head([winners]) } )
}
```

„Earlier this year, Eminem won two Grammy awards, one for his debut album The Slim Shady LP, and another for best rap solo performance on the single My Name Is.“

Abbildung 4: Regelbeispiel.

tierung allein auf Satzebene arbeitet. Es ist allerdings bekannt, dass Ereignisinstanzen oft über mehrere Sätze innerhalb eines Dokuments verteilt sind. Nachrichtentext, der sich mit Musikpreisen befasst, hat darüberhinaus zwei Eigenschaften, die man sich zunutze machen kann, um die Stelligkeit der extrahierten Instanzen zu erhöhen. Eine dieser Eigenschaften ist, dass die Artikel häufig sehr zeitnah mit der Preisverleihung berichten. Es lässt sich also mit einer bestimmten Gewissheit das Veröffentlichungsjahr eines Artikels als Argument für die darin berichteten Instanzen verwenden. Tabellen 15 und 16 zeigen die Ergebnisse bei diesem Vorgehen. Es fällt auf, dass auf diese Weise deutlich mehr höherstufige Instanzen extrahiert werden können. Es können sogar insgesamt mehr Instanzen extrahiert werden, da bei den vorigen Ergebnissen mehrere Vorkommen zweistelliger Instanzen vereinigt wurden, die aus Artikeln mit unterschiedlichen Erscheinungsjahren stammen. Es fällt jedoch ebenfalls auf, dass die Korrektheit der Instanzen deutlich sinkt. Dies mag damit zusammenhängen, dass das Jahr einer Musikpreisverleihung (wie etwa in *„the 2006 MTV Video Music Awards“*) nicht notwendigerweise das Jahr ist, in dem die Verleihungszeremonie stattfindet, sondern das Jahr, das für die Ehrung ausschlaggebend ist. Ein zukünftiges System könnte eventuell für jede Instanz eine Klassifikation auf der Basis bestimmter Merkmale der Textquellen, aus der sie stammt, vornehmen, um zu entscheiden, ob das Erscheinungsjahr des Textes als Argument der Instanz hinzugefügt werden soll oder nicht.

Stelligkeit	*extrahierte Instanzen*	*davon im Ideal Table*	*Abdeckung des Ideal Table*[6]
2	9	$66,67\%$	$11,90\%$
3	11	$36,36\%$	$16,33\%$
4	1	$100,00\%$	$20,00\%$

Tabelle 15: Ergebnisse nach Einbeziehung des Jahres der Textquelle als Argument (Instanzen als Seed).

Eine weitere Quelle an Information für die Vervollständigung niedrigstelliger Instanzen sind die Dokumenttitel. Die Titel von $12,87\%$ aller Dokumente enthalten die Erwähnung eines Künstlers oder Preises.

```
{ head(("have", V))
    -> subj( { head([winners])
                 -> rel( { head(("fin", C))
                              -> i( { head(("be", VBE))
                                        -> pred( { head([winners]) } )
    } ) } ) } )
    -> obj( { head(("nomination", N))
                -> nn( { head([prize]) } )
                -> mod( { head(("for", Prep))
                              -> pcomp-n( { head("Marshall Mathers LP")
                                              -> lex-mod( { head([winners]) } )
    } ) } ) } )
    -> mod_2( { head(("vpsc", C))
                  -> i_2( { head(("include", V))
                                -> conj( { head([area]) } )
    } ) } )
}
```

„Eminem, whose real name is Marshall Mathers III, has three Grammy nominations for The Marshall Mathers LP, including album of the year, best rap album, and best rap solo performance.“

Abbildung 5: Regelbeispiel.

Stelligkeit	*extrahierte Instanzen*	*davon im Ideal Table*	*Abdeckung des Ideal Table*[6]
2	9	$55,56\%$	$22,02\%$
3	29	$44,83\%$	$34,69\%$
4	2	$50,00\%$	$20,00\%$

Tabelle 16: Ergebnisse nach Einbeziehung des Jahres der Textquelle als Argument (Regeln als Seed).

6 Zusammenfassung

In dieser Arbeit wurde die Portierung eines Relations– und Ereignisextraktionssystems vorgestellt. Die Portierung beinhaltet zum einen die Adaption des bestehenden Extraktionssystems an die Domäne der Musikpreisverleihungen und zum anderen die Übertragung der Extraktionsgrammatik aus der Domäne der Nobelpreisverleihungen.

Der mit der Adaption des bestehenden Systems verbundene Aufwand war im Bereich der Named Entity–Erkennung am größten. Diese ist für die meisten in Abschnitt 2 vorgestellten Ansätze notwendig, und stellt einen Aufwand dar, der größtenteils unabhängig vom gewählten Verfahren getrieben werden muss. Bei den übrigen Anpassungen handelte es sich vorwiegend

um Feineinstellungen des Algorithmus.

Es wurde gezeigt, dass durch die Übertragung bereits bestehender Regeln aus einer Domäne mit hilfreichen Texteigenschaften deutlich mehr Instanzen extrahiert werden können als mit der alleinigen Verwendung von Instanzseeds. Es bleibt allerdings in beiden Fällen bei einer verbesserungswürdigen Abdeckung des Ideal Table. Die vorliegende Arbeit enthält einige Vorschläge, wie diese Werte in der Zukunft möglicherweise erhöht werden können.

A Kompletter Ideal Table

Künstler	*Jahr*	*Preis*	*Gebiet*
madonna	1986	golden_raspberry_award	worst_actress
madonna	1987	american_music_award	favorite_poprock_female_video_artist
madonna	1987	golden_raspberry_award	worst_actress
madonna	1987	mtv_video_music_award	best_female_video
madonna	1988	golden_raspberry_award	worst_actress
madonna	1989	billboard_award	best_video
madonna	1989	mtv_video_music_award	best_art_direction
madonna	1989	mtv_video_music_award	best_cinematography
madonna	1989	mtv_video_music_award	best_direction
madonna	1989	mtv_video_music_award	viewers_choice
madonna	1990	mtv_video_music_award	best_cinematography
madonna	1990	mtv_video_music_award	best_direction
madonna	1990	mtv_video_music_award	best_editing
madonna	1991	academy_award	best_original_song
madonna	1991	american_music_award	favorite_dance_single
madonna	1992	grammy_award	best_music_video_long_form
madonna	1993	golden_raspberry_award	worst_actress
madonna	1993	mtv_video_music_award	best_art_direction
madonna	1993	mtv_video_music_award	best_cinematography
madonna	1994	golden_raspberry_award	worst_actress
madonna	1995	mtv_video_music_award	best_female_video
madonna	1996	billboard_award	artist_achievement
madonna	1996	golden_raspberry_award	worst_supporting_actress
madonna	1997	academy_award	best_original_song
madonna	1997	golden_globe_award	best_actress
madonna	1997	golden_globe_award	best_original_song
madonna	1997	golden_globe_award	best_actress
madonna	1997	golden_globe_award	best_picture
madonna	1998	mtv_european_music_award	best_album
madonna	1998	mtv_european_music_award	best_female_artist
madonna	1998	mtv_video_music_award	best_choreography
madonna	1998	mtv_video_music_award	best_direction
madonna	1998	mtv_video_music_award	best_editing
madonna	1998	mtv_video_music_award	best_female_video
madonna	1998	mtv_video_music_award	best_special_effects
madonna	1998	mtv_video_music_award	best_video
madonna	1999	grammy_award	best_dance
madonna	1999	grammy_award	best_music_video_short_form
madonna	1999	grammy_award	best_pop_album
madonna	1999	mtv_video_music_award	best_video_from_a_film

madonna	2000	billboard_award	best_pop
madonna	2000	billboard_award	best_video
madonna	2000	golden_raspberry_award	worst_actress_of_the_century
madonna	2000	grammy_award	best_original_song
madonna	2000	mtv_european_music_award	best_dance
madonna	2000	mtv_european_music_award	best_female_artist
madonna	2001	brit_award	best_international_female
madonna	2001	golden_raspberry_award	worst_actress
madonna	2002	golden_raspberry_award	worst_actress
madonna	2002	golden_raspberry_award	worst_screen_couple
madonna	2002	golden_raspberry_award	worst_supporting_actress
madonna	2003	american_music_award	best_international_artist
madonna	2003	golden_raspberry_award	worst_actress
madonna	2003	golden_raspberry_award	worst_screen_couple
madonna	2003	golden_raspberry_award	worst_supporting_actress
madonna	2006	brit_award	best_international_female
madonna	2007	grammy_award	best_dance
eminem	1999	mtv_european_music_award	best_hiphop
eminem	1999	mtv_video_music_award	best_new_artist
eminem	2000	billboard_award	best_maximum_vision_video
eminem	2000	billboard_award	best_raphip_hop_clip
eminem	2000	grammy_award	best_rap_album
eminem	2000	grammy_award	best_rap_solo_performance
eminem	2000	mtv_european_music_award	best_album
eminem	2000	mtv_european_music_award	best_hiphop
eminem	2000	mtv_video_music_award	best_male_video
eminem	2000	mtv_video_music_award	best_rap_video
eminem	2000	mtv_video_music_award	best_video
eminem	2000	source_award	best_lyricist
eminem	2000	source_award	best_video
eminem	2001	billboard_award	best_rap_hip_hop_clip
eminem	2001	billboard_award	maximum_vision_award
eminem	2001	brit_award	best_international_male
eminem	2001	grammy_award	best_rap_album
eminem	2001	grammy_award	best_rap_performance
eminem	2001	grammy_award	best_rap_solo_performance
eminem	2001	meteor_ireland_music_award	best_international_single
eminem	2001	mtv_european_music_award	best_hiphop
eminem	2001	source_award	best_video
eminem	2002	billboard_award	best_album
eminem	2002	billboard_award	best_hiphop_album
eminem	2002	mtv_european_music_award	best_album
eminem	2002	mtv_european_music_award	best_hiphop
eminem	2002	mtv_european_music_award	best_male_artist

eminem	2002	mtv_video_music_award	best_direction
eminem	2002	mtv_video_music_award	best_male_video
eminem	2002	mtv_video_music_award	best_rap_video
eminem	2002	mtv_video_music_award	best_video
eminem	2003	academy_award	best_original_song
eminem	2003	american_music_award	favorite_album_hiphoprb
eminem	2003	american_music_award	favorite_album_poprock
eminem	2003	american_music_award	favorite_hiphoprb_male_artist
eminem	2003	american_music_award	favorite_poprock_male_artist
eminem	2003	billboard_award	best_album
eminem	2003	billboard_award	best_rap_album
eminem	2003	brit_award	best_international_album
eminem	2003	brit_award	best_international_male
eminem	2003	grammy_award	best_music_video_short_form
eminem	2003	grammy_award	best_rap_album
eminem	2003	meteor_ireland_music_award	best_international_male
eminem	2003	mtv_european_music_award	best_hiphop
eminem	2003	mtv_movie_award	best_actor
eminem	2003	mtv_movie_award	best_male_performance
eminem	2003	mtv_movie_award	breakthrough_male_performance
eminem	2003	mtv_movie_award	breakthrough_performance
eminem	2003	mtv_video_music_award	best_video_from_a_film
eminem	2004	grammy_award	best_rap_solo_performance
eminem	2004	grammy_award	best_rap_song
eminem	2004	mtv_european_music_award	best_hiphop
eminem	2005	american_music_award	favorite_hiphoprb_male_artist
eminem	2005	brit_award	best_international_male
eminem	2006	american_music_award	favorite_hiphoprb_male_artist
u2	1983	brit_award	best_live_performance
u2	1987	mtv_video_music_award	viewers_choice
u2	1988	brit_award	best_international_group
u2	1988	grammy_award	best_album
u2	1988	grammy_award	best_rock_performance
u2	1989	brit_award	best_international_group
u2	1989	grammy_award	best_rock_performance
u2	1989	mtv_video_music_award	best_video_from_a_film
u2	1990	brit_award	best_international_group
u2	1992	brit_award	best_international_group
u2	1992	mtv_video_music_award	best_group_video
u2	1992	mtv_video_music_award	best_special_effects
u2	1993	grammy_award	best_rock_performance
u2	1994	grammy_award	best_alternative_album
u2	1995	golden_globe_award	best_original_song
u2	1995	grammy_award	best_music_video_long_form

u2	1995	mtv_european_music_award	best_group
u2	1997	mtv_european_music_award	best_live_performance
u2	1998	brit_award	best_international_group
u2	2001	american_music_award	internet_artist_of_the_year
u2	2001	brit_award	best_international_group
u2	2001	brit_award	outstanding_contribution_to_music_award
u2	2001	grammy_award	best_rock_performance
u2	2001	grammy_award	record_of_the_year
u2	2001	grammy_award	song_of_the_year
u2	2001	meteor_ireland_music_award	best_irish_band
u2	2001	meteor_ireland_music_award	best_irish_rock_album
u2	2001	meteor_ireland_music_award	best_irish_songwriter
u2	2001	mtv_video_music_award	career_achievement
u2	2002	american_music_award	internet_artist_of_the_year
u2	2002	grammy_award	best_pop
u2	2002	grammy_award	best_rock_album
u2	2002	grammy_award	best_rock_performance
u2	2002	grammy_award	record_of_the_year
u2	2002	meteor_ireland_music_award	best_irish_live_performance
u2	2002	meteor_ireland_music_award	best_irish_rock_album
u2	2002	meteor_ireland_music_award	best_irish_rock_band
u2	2002	meteor_ireland_music_award	best_irish_rock_single
u2	2002	meteor_ireland_music_award	best_irish_songwriter
u2	2002	meteor_ireland_music_award	best_irish_video
u2	2003	golden_globe_award	best_original_song
u2	2003	meteor_ireland_music_award	best_group
u2	2003	meteor_ireland_music_award	humanitarian_award
u2	2004	golden_globe_award	best_original_song
u2	2005	grammy_award	best_music_video_short_form
u2	2005	grammy_award	best_rock_performance
u2	2005	grammy_award	best_rock_song
u2	2006	grammy_award	best_album
u2	2006	grammy_award	best_rock_album
u2	2006	grammy_award	best_rock_performance
u2	2006	grammy_award	best_rock_song
u2	2006	grammy_award	song_of_the_year
u2	2006	meteor_ireland_music_award	best_irish_album
u2	2006	meteor_ireland_music_award	best_irish_band
u2	2006	meteor_ireland_music_award	best_live_performance

B Extraktionsergebnisse

B.1 Instanzen als Seed

Künstler	*Jahr*	*Preis*	*Gebiet*
eminem	2001	grammy_award	⊥
eminem	⊥	grammy_award	best_rap_album
eminem	1997	⊥	⊥
eminem	1999	⊥	⊥
eminem	2000	⊥	⊥
eminem	2001	⊥	⊥
madonna	1985	⊥	⊥
madonna	1987	⊥	⊥
madonna	1991	⊥	⊥
u2	1978	⊥	⊥
u2	1987	⊥	⊥
u2	1992	⊥	⊥
eminem	⊥	grammy_award	⊥
eminem	⊥	mtv_video_music_award	⊥
madonna	⊥	grammy_award	⊥
madonna	⊥	mtv_video_music_award	⊥
u2	⊥	academy_award	⊥
u2	⊥	brit_award	⊥
u2	⊥	grammy_award	⊥

B.2 Regeln als Seed

Künstler	*Jahr*	*Preis*	*Gebiet*
eminem	⊥	academy_award	best_original_song
eminem	⊥	grammy_award	best_album
madonna	1985	⊥	⊥
madonna	1987	⊥	⊥
madonna	1991	⊥	⊥
eminem	1997	⊥	⊥
eminem	1999	⊥	⊥
eminem	2000	⊥	⊥
eminem	2001	⊥	⊥
u2	1978	⊥	⊥
u2	1987	⊥	⊥
u2	1992	⊥	⊥
madonna	⊥	golden_globe_award	⊥
madonna	⊥	grammy_award	⊥
madonna	⊥	mtv_video_music_award	⊥

eminem	⊥	academy_award	⊥
eminem	⊥	grammy_award	⊥
eminem	⊥	mtv_video_music_award	⊥
u2	⊥	academy_award	⊥
u2	⊥	brit_award	⊥
u2	⊥	grammy_award	⊥
madonna	⊥	⊥	best_female_artist
madonna	⊥	⊥	best_international_female
madonna	⊥	⊥	record_of_the_year
madonna	⊥	⊥	sexiest_woman
eminem	⊥	⊥	best_album
eminem	⊥	⊥	best_hiphop
eminem	⊥	⊥	best_male_artist
eminem	⊥	⊥	best_rap_album
eminem	⊥	⊥	best_song
eminem	⊥	⊥	best_video_from_a_film
eminem	⊥	⊥	breakthrough_male_performance
eminem	⊥	⊥	rap_artist
u2	⊥	⊥	best_group
u2	⊥	⊥	best_irish_band
u2	⊥	⊥	best_live_performance
u2	⊥	⊥	best_music_dvd
u2	⊥	⊥	outstanding_contribution_to_music_award

Literatur

[Agichtein und Gravano 2000] Agichtein, Eugene ; Gravano, Luis: SNOWBALL: Extracting relations from large plain-text collections. In: *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000

[Blaschke und Valencia 2002] Blaschke, Christian ; Valencia, Alfonso: The Frame-Based Module of the SUISEKI Information Extraction System. In: *IEEE Intelligent Systems* 17 (2002), Nr. 2, S. 14–20

[Brin 1999] Brin, Sergey: Extracting Patterns and Relations from the World Wide Web. In: *WebDB '98: Selected papers from the International Workshop on The World Wide Web and Databases*. London, UK : Springer, 1999, S. 172–183

[Bunescu und Mooney 2007] Bunescu, Razvan C. ; Mooney, Raymond J.: Extracting Relations from Text – From Word Sequences to Dependency Paths. In: Kao, Anne (Hrsg.) ; Poteet, Steve (Hrsg.): *Text Mining and Natural Language Processing*. Springer, 2007, S. 29–44

[Collins und Duffy 2001] Collins, Michael ; Duffy, Nigel: Convolution Kernels for Natural Language. In: *Proceedings of NIPS 2001*, 2001

[Doddington u. a. 2004] Doddington, George ; Mitchell, Alexis ; Przybocki, Mark ; Ramshaw, Lance ; Strassel, Stephanie ; Weischedel, Ralph: Automatic Content Extraction (ACE) program - task definitions and performance measures. In: *LREC 2004: Fourth International Conference on Language Resources and Evaluation*, 2004

[Drozdzynski u. a. 2004] Drozdzynski, Witold ; Krieger, Hans-Ulrich ; Piskorski, Jakub ; Schäfer, Ulrich ; Xu, Feiyu: Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. In: *Künstliche Intelligenz* 1 (2004), S. 17–23. – URL `http://www.kuenstliche-intelligenz.de/archiv/2004_1/sprout-web.pdf`

[Hong 2006] Hong, Li: *Automatische Ereignis- und Relationserkennung mit Seeds von variierender Komplexität*, Diplomarbeit, Universität des Saarlandes, Naturwissenschaftlich-Technische Fakultät I, Fachrichtung Informatik, Diplomarbeit, 2006

[Lin 1998] Lin, D.: Dependency-Based Evaluation of MINIPAR. In: *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation, Granada Spain*, 1998

[McCallum 2005] McCallum, Andrew: Information Extraction: distilling structured data from unstructured text. In: *ACM Queue* 3 (2005), Nr. 9, S. 48–57

[McDonald u. a. 2005] McDonald, Ryan T. ; Pereira, Fernando ; Kulick, Seth ; Winters, Scott ; Jin, Yang ; White, Peter S.: Simple Algorithms for Complex Relation Extraction with Applications to Biomedical IE. In: *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, 2005, S. 491–498

[Miller u. a. 2000] Miller, Scott ; Fox, Heidi ; Ramshaw, Lance ; Weischedel, Ralph: A Novel Use of Statistical Parsing to Extract Information from Text. In: *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2000, S. 226–233

[Mooney und Bunescu 2005] Mooney, Raymond J. ; Bunescu, Razvan: Mining Knowledge from Text Using Information Extraction. In: *SIGKDD Explorations Newsletter* 7 (2005), Nr. 1, S. 3–10

[Riloff 1996] Riloff, Ellen: Automatically Generating Extraction Patterns from Untagged Text. In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, MIT Press, 1996, S. 1044–1049

[Roth und Yih 2002] Roth, Dan ; Yih, Wen-tau: Probabilistic reasoning for entity & relation recognition. In: *Proceedings of the 19th international conference on Computational linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, 2002, S. 1–7

[Xu u. a. 2006] Xu, Feiyu ; Uszkoreit, Hans ; Li, Hong: Automatic Event and Relation Detection with Seeds of Varying Complexity. In: *Proceedings of AAAI 2006 Workshop Event Extraction and Synthesis*, 2006

[Xu u. a. 2007] Xu, Feiyu ; Uszkoreit, Hans ; Li, Hong: A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic : Association for Computational Linguistics, June 2007, S. 584–591. – URL `http://www.aclweb.org/anthology/P/P07/P07-1074`

[Zelenko u. a. 2003] Zelenko, Dmitry ; Aone, Chinatsu ; Richardella, Anthony: Kernel methods for relation extraction. In: *Journal of Machine Learning Research* 3 (2003), S. 1083–1106

[Zhao und Grishman 2005] Zhao, Shubin ; Grishman, Ralph: Extracting relations with integrated information using kernel methods. In: *ACL '05:*

Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown, NJ, USA : Association for Computational Linguistics, 2005, S. 419–426