Warum „Named Entities“ für die Chunk-Analyse wichtig sind*

*Ilona Steiner***

Zusammenfassung

In diesem Papier wird untersucht, inwiefern in Systemen zur syntaktischen Annotation von deutschen Korpora die Erkennung von „Named Entities“ einen notwendigen Vorverarbeitungsschritt für eine nachfolgende Chunk-Analyse darstellt. Es wird gezeigt, welche Arten von vorstrukturierten „Named Entity“-Ausdrücken einer Chunk-Analyse nutzen und welche Probleme sich andernfalls ergeben würden. Weiterhin wird ein Modul zur Erkennung dieser „Named Entities“ in deutschen XML-Dokumenten vorgestellt. Es ist so konzipiert, daß es in Systemen zur automatischen Annotation von Korpora einer Chunk-Erkennung vorgeschaltet werden kann.

23.1. Einleitung

Die Erkennung von sogenannten „Named Entities“ (z. B. Eigennamen, numerische Ausdrücke, Datumsangaben, etc.) wird im Bereich der automatischen Verarbeitung von freien Texten zunehmend wichtiger. Die Klassifikation dieser Kategorien bietet einerseits die Grundlage für (intelligentes) Suchen auf Korpora und stellt andererseits einen wichtigen Vorverarbeitungsschritt für komplexere Aufgaben dar.

Im Bereich der Informationserschließung bilden Systeme zur Erkennung von „Named Entities“ (NEs) die Basis für komplexere Aufgaben. Für das Englische gibt es mehrere Systeme (Mikheev et al., 1998, Borthwick et al., 1998, Black et al., 1998), über Systeme zum Deutschen gibt es allerdings kaum Publikationen (z. B. Piskorski und Neumann, 2000). Die verwendeten Annotationskriterien in diesen Systemen (siehe z. B. Chinchor, 1998, Piskorski und Neumann, 2000) umfassen teilweise recht komplexe NE-Konstruktionen (z. B. *von fast 30 auf über 50 Mio. Dollar*) und sind für Aufgaben zur Informationsextraktion ausgelegt.

In diesem Papier soll der Frage nachgegangen werden, ob in Systemen zur syntaktischen Annotation von deutschen Korpora eine „Named Entity“-Erkennung ebenfalls einen notwendigen Vorverarbeitungsschritt für eine nachfolgende Chunk-Analyse[1] darstellt. Welche Arten von vor-

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 245–252. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`

** Herzlich bedanken möchte ich mich bei Kathrin Beck, Badreddin Abolmaali und Ralph Albrecht für die Evaluation des Testkorpus. Für wertvolle Kommentare und Diskussionen danke ich Sandra Kübler, Manfred Sailer, Janina Radó, Ralph Albrecht, Heike Telljohann und Valia Kordoni.

[1] Bei einer Chunk-Analyse handelt sich um einen Ansatz zum partiellen Parsing syntaktischer Strukturen, der von Steven Abney entwickelt wurde (siehe Abney, 1991, 1996).

strukturierten NE-Ausdrücken nutzen der Chunk-Analyse bzw. welche Probleme würden sich andernfalls ergeben?

Anschließend wird das Modul NENA (**N**amed **EN**tity **A**nnotation) zur Erkennung von „Named Entities“ im Deutschen vorgestellt. NENA ist so konzipiert, daß es in Systemen zur automatischen Annotation von Korpora einer Chunk-Erkennung vorgeschaltet werden kann. Hierbei darf die NE-Annotation Chunk- oder Phrasengrenzen nicht überschreiten, was in den oben erwähnten Systemen jedoch der Fall ist.

23.2. Probleme beim Chunken nicht-annotierter „Named Entities“

Um die Notwendigkeit einer NE-Vorverarbeitung (z. B. durch NENA) zu demonstrieren, soll im folgenden gezeigt werden, welche Probleme entstehen, wenn ein Chunk-Parser NE-Ausdrücke analysieren soll, die nicht bereits vorstrukturiert sind. Dabei wurden folgende NE-Kategorien berücksichtigt: Temporale Ausdrücke (Datum, Uhrzeit), Eigennamen (Personen), numerische Ausdrücke, quantitative Ausdrücke (Maßangaben, Währungs- und Prozentangaben). Die Untersuchung ergab vier verschiedene Arten von Problemfällen:

1. **Komplexe Prämodifikation (durch Zahlen)**

   Komplexe numerische Ausdrücke mit folgender Zielstruktur:

   (1) [[1 Komma 2] Mio] Dollar
   CARD NN CARD NN NN

   würden mit einer reinen Chunk-Analyse fälschlicherweise folgende Struktur zugeordnet bekommen:

   (2) [1 Komma] [2 Mio] Dollar
   CARD NN CARD NN NN

   Für die Erkennung von komplexen numerischen Ausdrücken muß die lexikalische Ebene berücksichtigt werden (evtl. kombiniert mit Wortarten[2]). Die alleinige Berücksichtigung von Wortarten, wie es beim Chunking im allgemeinen üblich ist, führt zu Übergenerierung.

2. **Postmodifikation (durch Zahlen)**

   Eine Datumsangabe mit folgender Zielstruktur:

   (3) am [3. Januar 1967]
   CARD NN CARD

   würde fälschlicherweise folgende Struktur erhalten:

   (4) am [3. Januar] [1967]
   CARD NN CARD

[2] In diesem Papier und in NENA wird das Stuttgart–Tübingen Tagset (STTS) verwendet (wie in Schiller et al., 1995, beschrieben).

Die Jahreszahl als Postmodifikator würde bei einer reinen Chunk-Analyse wie in (4) nicht in den Datumsausdruck mit eingebunden werden, mit entsprechenden Auswirkungen für die Weiterverarbeitung. Das gleiche strukturelle Problem ergibt sich generell, wenn ein postnominaler Zahlausdruck mit eingebunden werden muß, wie z. B. in Uhrzeitangaben (*15 Uhr 30*) oder Währungsangaben (*10 Mark 50*).

3. **„Koordination" von Maßeinheit und Untereinheit(en)**

   Eine Währungsangabe mit der Zielstruktur:

   (5) [6 Mark und 50 Pfennige]
   CARD NN KON CARD NN

   würde mit einer reinen Chunk-Analyse folgende Struktur erhalten:

   (6) [6 Mark] und [50 Pfennige]
   CARD NN KON CARD NN

   Der Bezug zwischen der Einheit und der Untereinheit kann in (6) nicht hergestellt werden. Hier wird spezielle Information benötigt, welches die Einheiten bzw. Untereinheiten sind, die außerdem in einer spezifischen Reihenfolge angeordnet sind. Das gleiche Problem ergibt sich auch für Maßangaben (z. B. *2 Meter und 50 Zentimeter*).

4. **Mehrgliedrige Eigennamen**

   Eigennamen-Ausdrücke mit folgender Zielstruktur:

   (7) [John F. Kennedy]
   NE NE NE

   würden vermutlich (je nach Verarbeitungsstrategie) folgende Struktur erhalten:

   (8) [John F.] [Kennedy]
   NE NE NE

   Die Wortarten alleine sind nicht ausreichend, um die korrekte Struktur zu erhalten. Es wird spezielle Information benötigt, welches ein Vorname und welches ein Nachname ist. Dann können Regeln für deren Kombination spezifiziert werden.

Zusammenfassend kann gesagt werden, daß für die „Named Entity"-Kategorien numerische Ausdrücke, Eigennamen, Datums- und Uhrzeitangaben, Maß- und Währungsangaben eigene Subgrammatiken notwendig sind. Es handelt sich hierbei um spezielle Regeln, die auch die lexikalische Ebene berücksichtigen. Für die Kategorie Prozentangaben ist dies nicht unbedingt notwendig, da die Einheit Prozent keine Untereinheiten besitzt und in sich nicht komplexer ist als ein einfaches Nomen.

Generell ist es sinnvoll, eine Vorstrukturierung der NE-Ausdrücke in einem separaten Modul vor der Chunk-Analyse vorzunehmen. Die erkannten Strukturen können dann beim Chunking als Einheit weiterverarbeitet werden.

23.3. Das Modul NENA

NENA soll eine Vorstrukturierung der oben erwähnten „Named Entity“-Kategorien leisten und einer Chunk-Analyse vorgeschaltet werden können. Das Modul ist integrierbar in Systeme zur automatischen Annotation von Korpora und kann bei Bedarf erweitert werden.

23.3.1. Annotationsprinzipien

Um unerwünschte Interaktionen der NE-Annotationsebene mit der Chunk-Ebene zu vermeiden und die Modularität von NENA zu gewährleisten, sind folgende Annotationsprinzipien definiert worden:

- NENA strukturiert „Named Entities“, die aus mehreren Token bestehen. Aus einem Token bestehende „Named Entities“ (wie z.B. *Februar*, ohne Angabe des Tages oder Jahres) werden bei Bedarf markiert, jedoch strukturell nicht berührt.
- Strukturell zusammengefügt werden ausschließlich „Named Entities“, die keine Chunk- oder Phrasengrenzen überschreiten. D. h. es werden nur „reine“ NEs annotiert und keine komplexen Chunks oder Phrasen, die „Named Entities“ enthalten. Die komplexe Präpositionalphrase *von fast 30 auf über [[50 Mio.] Dollar]* wird folglich nur lokal geklammert.
- Die „Named Entity“-Annotation verläßt die Subgrammatik nicht. D. h. Artikel oder Modifikatoren, die sich auf eine „Named Entity“ beziehen, werden nicht mit eingebunden (z. B. *der schöne [3. Januar 1976]*). Dies ist die Aufgabe der Chunk-Grammatik.

23.3.2. Aufbau des Moduls

NENA erwartet als Eingabe einen tokenisierten und PoS-getaggten deutschen Text im XML-Format. Daraufhin werden mehrere modulare Grammatiken (in der Reihenfolge: numerische Ausdrücke, quantitative Ausdrücke, Datum, Uhrzeit, Eigennamen) nacheinander abgearbeitet. Dabei kann die Information der jeweils vorhergehenden Stufe genutzt werden. Ausgegeben wird das XML-Dokument, angereichert mit NE-Strukturen und NE-Attributen. Für „Named Entity“-Ausdrücke wurde das Element „NE“ eingeführt mit einem Attribut „T“ (Type) für die NE-Kategorie und einem optionalen Attribut „TS“ (Subtype) für die jeweilige Unterkategorie. Eine Erweiterung des Moduls um zusätzliche Kategorien oder Unterkategorien läßt sich somit problemlos über die Definition neuer Werte für die Attribute T bzw. TS erreichen. Nachfolgend wird ein Beispielsatz (*Thomas H. gab 31 Millionen Dollar aus.*) in XML gezeigt, vor und nach der NE-Annotation (siehe 9 bzw. 10).

(9) NENA-Eingabe:

```
<S><W PT="NE">Thomas</W><W PT="NE">H.</W>
<W PT="VVFIN">gab</W><W PT="CARD">31</W>
<W PT="NN">Millionen</W><W PT="NN">Dollar</W>
<W PT="PTKVZ">aus</W><W PT="$">.</W></S>
```

(10) NENA-Ausgabe:

```
<S><NE T="NAME"><W PT="NE">Thomas</W>
<W PT="NE">H.</W></NE><W PT="VVFIN">gab</W>
```

```
<NE T="QUANTITY" TS="CURRENCY">
<NE T="NUMBER"><W PT="CARD">31 </W>
<W PT="NN">Millionen</W></NE>
<W PT="NN">Dollar</W></NE><W PT="PTKVZ">aus</W>
<W PT="$.">.</W></S>
```

In (10) sind die NE-Kategorien NAME, NUMBER und QUANTITY (als Werte des Attributs T) annotiert. QUANTITY ist durch die Unterkategorie CURRENCY (als Wert des Attributs TS) nochmals genauer spezifiziert.

Für die Abarbeitung der Grammatiken wurde das Tool „fsgmatch“ der Language Technology Group (LTG) Edinburgh (Grover et al., 1999) benutzt, das es erlaubt, mit Hilfe mehrerer hintereinandergeschalteter Finite-State-Transducers XML-Dokumente anzureichern. Der bereitgestellte Regelformalismus nutzt Pattern-Matching-Techniken und erlaubt die Benutzung von regulären Ausdrücken, Lexika und die Einbeziehung des Kontexts.

23.3.3. Die „Named Entity“-Kategorien

Numerische Ausdrücke (`T="NUMBER"`):

Mit dieser Subgrammatik werden komplexe numerische Ausdrücke, die aus mehreren Token bestehen, strukturell zusammengefaßt, wie z. B. *[31 Millionen] Zuschauer*, *[Wurzel zwei]*, *[18 mal 10E minus 6] Kilowattstunden*, usw. Die Grammatik benutzt dafür lexikalische Elemente und Wortarten. In NENA werden diese Ausdrücke abgegrenzt gegenüber Rechenoperationen wie *10 plus 10* oder Formatangaben wie *2 mal 2 Meter*. Hierfür könnten eigene Unterkategorien geschaffen werden. Von einer nachfolgenden Chunk-Ebene kann der komplexe numerische Ausdruck wie eine einfache Kardinalzahl weiterverarbeitet werden.

Quantitative Ausdrücke (`T="QUANTITY"`):

Es werden die folgenden Unterkategorien unterschieden:

- **Maßangaben** (`TS="MEASURE"`):

  Maßangaben, die aus mehreren Token bestehen, werden strukturell zusammengefaßt, wie z. B. *[einen Meter (und) zwanzig (Zentimeter)]*, *[1 Pfund] Kirschen*, *[minus 32 Grad Celsius]*. Für die Erkennung der „koordinierten“ Strukturen (wie z. B. *[2m, 50 cm und 4 mm]*) wird den Maßeinheiten im Lexikon eine Rangordnung der Einheiten über Tags zugewiesen. Über Regeln wird spezifiziert, daß pro Einheit bzw. Untereinheit nur ein Konjunkt zugelassen ist und die Konjunkte in einer spezifischen Reihenfolge stehen müssen. Von einer nachfolgenden Chunk-Ebene kann die Maßangabe wie ein einfaches Nomen weiterverarbeitet werden.

- **Währungsangaben** (`TS="CURRENCY"`):

  Strukturell zusammengefaßt werden Währungsangaben, die aus mehreren Token bestehen, wie z. B. *[drei Mark (und) fünfzig (Pfennige)]*. Bei der Erkennung der „koordinierten“ Strukturen wird das gleiche Prinzip verwendet, wie bereits oben unter Maßangaben erläutert. Die Währungsangabe kann von einer nachfolgenden Chunk-Ebene wie ein einfaches Nomen

weiterverarbeitet werden. D. h. sie kann anschließend mit einem Artikel oder mit Modifikatoren zu einem Nominalchunk zusammengefügt werden (z. B. *die versprochenen [3 Mark 50]*).

- **Prozentangaben** (`TS="PERCENTAGE"`)**:**

  Prozentangaben werden strukturell zusammengefaßt, wie z. B. *[[eins Komma zwei] Prozent]*. Eine solche Angabe kann von einer nachfolgenden Chunk-Ebene wie ein einfaches Nomen weiterverarbeitet werden.

Die drei Unterkategorien der quantitativen Ausdrücke werden mit einer gemeinsamen Grammatik abgedeckt, die die lexikalische Ebene, Wortarten, ein Lexikon sowie die Vorstrukturierung der numerischen Ausdrücke verwendet. Beispiele für die Einbettung von erkannten numerischen Ausdrücken in quantitative Ausdrücke sind *[[31 Mio.] Quadratkilometer/Mark]* bzw. *[[neunzig Komma neun] Prozent]*.

Temporale Ausdrücke (`T="TEMPORAL"`):

Es werden die folgenden Unterkategorien unterschieden:

- **Datum** (`TS="DATE"`):

  Datumsangaben, die aus mehreren Token bestehen, werden strukturell zusammengefaßt, wie z. B. *der [dritte August siebenundsechzig]*, *[25. IV.]*, *im [Herbst 1911 n. Chr.]*, etc. Die Datumsausdrücke werden mit Hilfe zweier Grammatiken erkannt: eine Grammatik erkennt, ob es sich bei einem Token um die Angabe eines Tages, eines Monats oder eines Jahres handelt. Diese Information wird den einzelnen Token über ein zusätzliches Attribut „NE“ des Elements „W“ hinzugefügt (siehe 11). Eine zweite Grammatik, die anschließend durchlaufen wird, fügt die erkannten Token (evtl. plus weitere Angaben) dann strukturell zu einem Datumsausdruck zusammen. In beiden Grammatiken wird neben der lexikalischen Ebene und den Wortarten ein Lexikon verwendet.

  Es handelt sich dabei um eine Modularisierung innerhalb der Datumsangaben. Die strukturelle Ebene des Datumsausdrucks wurde von der lexikalischen Ebene getrennt, um komplexe Interaktionen der beiden Ebenen zu vermeiden. D. h. die Regeln für die verschiedenen Kombinationsmöglichkeiten der Einheiten (Tag, Monat, Jahr, Jahreszeit, Bezug zu Christi Geburt, usw.) innerhalb der Datumsstruktur wurden getrennt von den Regeln für die konkreten Erscheinungsformen dieser Einheiten im Text (die Monatsangabe kann z. B. ausgeschrieben (evtl. abgekürzt), in arabischen oder in römischen Zahlen vorkommen). (11) ist ein Beispiel für einen NE-annotierten Datumsausdruck (*6. Okt. 1939*) im XML-Format nach dem Durchlaufen beider Grammatiken:

  (11)
```
<NE T="TEMPORAL" TS="DATE"><W NE="DAY" PT="ADJA">6.</W>
<W NE="MONTH" PT="NN">Okt.</W>
<W NE="YEAR" PT="CARD">1939</W></NE>
```

  Von einer nachfolgenden Chunk-Ebene kann der Datumsausdruck wie ein einfaches Nomen weiterverarbeitet werden.

NE-Kategorie	Precision	Recall	F-Measure	Häufigkeit
NUMBER	100%	99,4%	99,7	170
MEASURE	100%	94,5%	97,1	164
CURRENCY	100%	97,9%	98,9	94
PERCENTAGE	100%	98,8%	99,3	81
DATE	100%	99,2%	99,6	259
TIME	100%	100%	100	45

Tabelle 23.1.: Evaluationsergebnisse

- **Uhrzeit** (`TS="TIME"`):

  Uhrzeitangaben, die aus mehreren Token bestehen, werden strukturell zusammengefaßt, wie z. B. *[13 Uhr 15]*, *[fünf vor zwölf]*, *um [halb acht]*, etc. Die Grammatik verwendet hierfür neben lexikalischen Elementen und Wortarten auch ein Lexikon. Von einer nachfolgenden Chunk-Ebene kann die Zeitangabe wie ein einfaches Nomen weiterverarbeitet werden. D. h. die Zeitangabe kann anschließend entweder direkt zu einem Nominalchunk erweitert oder mit prämodifizierenden Elementen zu einem Nominalchunk zusammengefaßt werden (z. B. *[ungefähr [15 Uhr 30]]*).

Eigennamen von Personen (`T="NAME", TS="PERSON"`):

Eigennamen, die aus mehreren Token bestehen, werden strukturell zusammengefaßt, wie beispielsweise *[Thomas H.]*, *[Franz v. Assisi]*, *[Henri IV.]*, etc. Die Grammatik verwendet neben lexikalischen Elementen und Wortarten ein Vornamen-Lexikon. Über Regeln wird spezifiziert, welche Kombinationen von Vornamen und Nicht-Vornamen, Adelszusätzen, etc. erlaubt sind. Von einer nachfolgenden Chunk-Ebene kann der mehrgliedrige Namensausdruck wie ein einfaches Nomen weiterverarbeitet werden.

23.4. Evaluierung

Das Modul NENA wurde auf Zeitungstexten getestet (Ausschnitte des Mannheimer Korpus I[3]) und mit den anfangs festgelegten Annotationskriterien verglichen. Insgesamt wurden ca. 62 000 Token durchsucht. Die Ergebnisse für Precision, Recall und F-Measure sowie die absolute Häufigkeit der NEs im Testkorpus sind in Tab. 23.1 für jede Unterkategorie dargestellt.[4]

Die geringe Fehlerquote ist teilweise auf das relativ kleine Testkorpus und der damit verbundenen kleinen Anzahl der NE-Vorkommen zurückzuführen. Andererseits ermöglichen die in Abschnitt 23.3.1 definierten Annotationsprinzipien auch eine sehr hohe Erkennungsrate.

[3] Es handelt sich beim Mannheimer Korpus I um ein repräsentatives, schriftsprachliches Korpus des Instituts für deutsche Sprache, Mannheim. Umfang: 2 570 970 Token

[4] Die Kategorie Eigennamen wurde nicht evaluiert, da die Grammatik momentan nur mit einem kleinen Lexikon implementiert ist.

Beispiele für nicht erkannte „Named Entities“ sind Ausdrücke mit seltenen Maßeinheiten, wie z. B. *Klafter*, *Gran*, bzw. ambigen Einheiten wie *Pfund*, *Drachmen* (Maßeinheit vs. Währungseinheit), *Fuß* oder *Stadien* (Maßeinheit vs. Sportarenen). Ein strukturelles Problem ergab sich beispielsweise bei der Währungsangabe *1,67 Milliarden (plus 86 Millionen) Mark*, wo der Bezug zur Währungseinheit aufgrund der Klammern im Text nicht hergestellt werden konnte.

23.5. Zusammenfassung

Es wurde gezeigt, daß es notwendig ist, vor der Chunk-Analyse eine Strukturierung von „Named Entities“ vorzunehmen, um eine fehlerhafte Analyse dieser Ausdrücke (und u. U. auch der weiterführenden syntaktischen Analyse) zu vermeiden. Dies gilt für alle untersuchten NE-Kategorien mit Ausnahme der Prozentangaben. Außerdem wurde das Modul NENA zur Erkennung von NEs vorgestellt, in dem die verschiedenen Subgrammatiken der untersuchten Kategorien modular implementiert sind. NENA ist so konzipiert, daß es keine unerwünschten Interaktionen mit einer nachfolgenden Chunk-Ebene zeigt.

Offen bleibt, ob die Ergebnisse auch für andere NE-Kategorien gültig sind, wie z. B. mathematische Formeln (*y gleich 2 hoch x*) oder Sportdisziplinen (*4 mal 200 m Delphin*), etc. Ein weiteres Problem stellt die Koordination (z. B. von „Named Entities“) dar. Hierfür müßte untersucht werden, welche Koordinations-Phänomene bereits auf der NE-Ebene und welche erst auf einer späteren Analyseebene erkannt werden können.

Literaturverzeichnis

Abney, S. (1991): “Parsing by Chunks”. In: *Principle-Based Parsing*, herausgegeben von Berwick, R.; Abney, S. und Tenney, C., Kluwer.

Abney, S. (1996): “Partial Parsing via Finite-State Cascades”. In: *Workshop on Robust Parsing (ESSLLI 1996)*, herausgegeben von Carroll, J.

Black, W. J.; Rinaldi, F. und Mowatt, D. (1998): “FACILE: Description of the NE System used for MUC-7”. In: *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Borthwick, A.; Sterling, J.; Agichtein, E. und Grishman, R. (1998): “NYU: Description of the MENE Named Entity System as used in MUC-7”. In: *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Chinchor, N. (1998): “MUC-7 Named Entity Task Definition”. In: *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Grover, C.; Matheson, C. und Mikheev, A. (1999): “TTT: Text Tokenisation Tool version 1.0”. Language Technology Group, University of Edinburgh. Online verfügbar: `http://www.ltg.ed.ac.uk/software/ttt/`.

Mikheev, A.; Grover, C. und Moens, M. (1998): “Description of the LTG System used for MUC-7”. In: *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Piskorski, J. und Neumann, G. (2000): “An Intelligent Text Extraction and Navigation System”. In: *6th International Conference on Computer-Assisted Information Retrieval (RIAO-2000)*.

Schiller, A.; Teufel, S. und Thielen, C. (1995): “Guidelines für das Tagging deutscher Textcorpora mit STTS”. Technischer Bericht, Universitäten Stuttgart und Tübingen. Online verfügbar: `http://www.sfs.nphil.uni-tuebingen.de/ELWIS/stts/stts.html`.