Exposé zur Diplomarbeit

EquatorNLP
Erkennen von Angaben über Erdbebenschäden in Textmeldungen

Lars Döhling

21. Februar 2010

Betreuer: Prof. Dr. Ulf Leser*

*HU-Berlin

1 Ziel

Im Rahmen der Arbeit soll untersucht werden, in wieweit es möglich ist, Angaben zu Schäden in Textmeldungen über Erdbeben automatisch zu erkennen. Dazu wird eine sprachverarbeitende Softwarekomponente erstellt, deren Funktionalität anschließend in ein bestehendes Informationsextraktionssystem eingebunden wird.

2 Motivation

Mehrere Male im Jahr bebt die Erde so stark, dass dabei Menschen zu Schaden kommen [U.S10a][U.S10b]. In der Folge sind Informationen wie das Ausmaß der Schäden und die Zahl der Verletzen wichtige Entscheidungsfaktoren über eventuelle Hilfseinsätze und deren Koordinierung. Da diese Daten meist dezentral und in Form von natürlichsprachlichen Texten vorliegen[1], gilt es zu untersuchen, in wieweit IT-Systeme diese Prozesse unterstützen können.

Mit einem gemeinsam mit der Deutsche TaskForce Erdbeben[2] entwickelten CMS – genannt **Equator** – wurde bereits ein solches entworfen und implementiert [Döh08]. **Equator** extrahiert einige der über ein frei wählbares Beben bei vier externen Quellen verfügbaren Daten, integriert diese und generiert eine gemeinsame Kartenansicht.

Wie für ein CMS charakteristisch, besteht die Möglichkeit manuell Texte im System zu hinterlegen. Diese Texte werden oftmals aus anderen Quellen – z.B. Zeitungen – kopiert. Daraus ergibt sich unmittelbar die zentrale Frage der Informationsextraktion (nach [Mit05, S. 545]):

> *Welches Wissen steckt in diesen Freitexten und mit welchen Methoden kann es automatisch erschlossen werden?*

Es soll nun in dieser Arbeit versucht werden, diese Frage zum Teil zu beantworten. Ziel ist die von **Equator** geleistete Informationsextraktion, welche bisher auf geologische Informationen

[1] Am Beispiel des Haiti-Bebens vom Januar 2010 wird die Bandbreite an Informationen sichtbar. Während das USGS nur eine Zusammenfassung der Schäden wiedergibt [U.S10c]

> *„[...]At least 112,405 people killed, 196,595 injured, 800,000 to 1 million displaced and severe to extreme damage in the Port-au-Prince area.[...]“,*

bieten die BBC News detailliert Angaben [BBC10]

> *„[...]The AFP news agency quoted the Jordanian army as saying three of its peacekeepers had been killed and 21 wounded.[...]“.*

[2] Sitz im GeoForschungsZentrum Potsdam, http://www.gfz-potsdam.de

zum Beben beschränkt ist, um das Erkennen von Angaben über Erdbebenschäden in Texten zu erweitern.

Für die Umsetzung soll ein Musterabgleich auf so genannten Dependenzgraphen zum Einsatz kommen. Beim zugrundeliegenden Dependenzmodell wird die syntaktische Struktur der Sätze in einem Text als eine Menge von Relationen zwischen den Wörtern der Sätze aufgefasst [CEE+04, S. 233]. Dies lässt sich als Graph veranschaulichen, bei dem die Wörter die Knoten und die Relationen die Kanten bilden.

In der Informationsextraktion wurde und wird versucht, unter Zuhilfenahme der syntaktischen Struktur eines Satzes auf die semantischen Beziehungen seiner Wörter zu schließen (Relationsextraktion, RE). Das Dependenzmodell eignet sich dafür potentiell gut, da hier - im Gegensatz zu sequenzbasierten Ansätzen wie reguläre Ausdrücke - leichter Relationen zwischen entfernt stehenden Wörtern eines Satzes erfasst werden können [FKZ07][CS07].

3 Vorgehen

Es lassen sich grob folgende Schritte unterscheiden

- Korpus erzeugen
- Sprachverarbeitende Softwarekomponente erstellen
- Evaluation
- Anbindung an Equator

3.1 Korpus

Um eine Aussage darüber treffen zu können, wie gut die gewählte Methode die Fragestellung beantworten kann, benötigt man als Vorlage eine vollständig und korrekt annotierte Textsammlung (das Korpus [LZ06, S. 7]). D.h. alle für die Fragestellung „interessanten“ Begriffe – genannt Entitäten – sowie deren Relationen sind markiert. Ein solches als Referenz für eine Evaluation (siehe 3.3) fungierendes Korpus wird als Gold Standard bezeichnet [LZ06, S. 143].

In der Arbeit wird auf drei Entitätstypen

- `Anzahl` - *„23“*, *„five“*, *„many“* ...
- `Anzahlmodifikator` - *„more then“*, *„at least“* ...

- `Schadenindikator` - *„killed“*, *„death“* ...

und eine Relation

- `Erdbebenschaden`(`Anzahlmodifikator`, `Anzahl`, `Schadenindikator`)

fokussiert.

Da kein geeignetes Korpus verfügbar ist, muss zunächst ein solches erstellt werden. Dazu werden englische Texte aus dem WWW manuell gesammelt, normalisiert, tokenisiert und annotiert. Die Sätze des Korpus werden anschließend mit Hilfe eines Parsers – wie z.B. des Charniak-Johnson-Parsers [CJ05] in Kombination[3] mit dem Stanford-Konverter [MMM06] – in Dependenzgraphen transformiert.

Das Korpus wird in eine Trainings- und Testmenge aufgeteilt. Für die Erstellung der sprachverarbeitenden Softwarekomponente wird ausschließlich die Trainingsmenge herangezogen.

3.2 Sprachverarbeitende Softwarekomponente

Grundsätzlich hat sich die Umsetzung als eine Kette von sprachverarbeitenden Softwarekomponenten (NLP-Pipeline[4]) bewährt. Die Ausgabe eines Verarbeitungsschrittes dient hierbei als Eingabe des Nächsten.

Ausgehend von einem mit Entitäten und Relationen annotierten Korpus lassen sich folgende Schritte unterscheiden

- Identifizieren der Entitäten
- Identifizieren der Relationen

3.2.1 Entitätserkennung

Um später in unbekannten Texten Entitätsrelationen erkennen zu können, ist es zunächst notwendig alle Entitäten zu identifizieren (NER[5]). Zu diesem Zweck werden für `Anzahlmodifikator` und `Schadenindikator` aus dem Trainingskorpus Lexika – angereichert mit Synonymen und manuellen Ergänzungen – generiert. Für `Anzahl` werden reguläre Ausdrücke verwendet.

[3] Der Charniak-Johnson-Parser generiert einen Konstituentenbaum. Diese zweite Art der syntaktischen Strukturierung von Sätzen in einem Text verwendet neben Wörtern (wie beim Dependenzmodell) auch komplexere Worteinheiten – genannt Konstituenten oder Phrasen. In einem Satz gibt es bei diesem Modell nicht nur Relationen zwischen Wörtern, sondern auch zwischen Konstituenten [CEE+04, S. 233]. Ein Konstituentenbaum lässt sich mit dem Stanford-Konverter in einen Dependenzgraphen umwandeln.

[4] von *natural language processing pipeline*

[5] von *named entity recognition*

Die NER-Komponente wird somit als eine auf Übereinstimmung mit dem Lexikon/regulären Ausdruck- basierte Klassifikation umgesetzt.

3.2.2 Relationsextraktion

Wie oben bereits erwähnt, sollen Relationen über Muster in Dependenzgraphen erkannt werden. Als Muster hat sich hierbei ein kürzester Pfad zwischen zwei in Relation stehenden Entitäten bewährt [BM05], als Algorithmus z.B. in JGraphT[6] enthalten. Da `Erdbebenschaden` eine trinäre Relation ist, wird ein Umweg über binäre Relationen begangen, aus denen sich im Anschluss die gesuchte höhere Relation synthetisieren lässt [MPK+05]. Analog zur Entitätserkennung wird ein Musterkatalog aus den im Trainingskorpus enthaltenen binären Relationen generiert.

In der RE-Komponente werden bei der Verarbeitung neuer Sätze die im Katalog vorhandenen Muster mit dem Dependenzgraphen abgeglichen. Gibt es eine Übereinstimmung, so werden die Entitäten als in Relation stehend klassifiziert, sonst nicht.

3.3 Evaluation

Nach Erstellung der NLP-Pipeline wird diese auf dem Testkorpus angewendet und die Trefferquote R[7], Genauigkeit P[8]und das kombinierte *F1*-Maß bestimmt [MS00, S. 267ff]. Dies geschieht sowohl unabhängig für die NER- und RE-Komponente, also auch für die Kombination der beiden.

Als Vergleich zum Dependenzmodell soll eine Methode dienen, welche die binären Relationen (siehe 3.2.2) anhand der Tokendistanz im Satz (hier als Liste von Tokens und nicht als Dependenzgraph) klassifiziert. Dabei wird die jeweils nächste passende Entität als in Relation stehend betrachtet. Die anschließende Rekonstruktion der Relation `Erdbebenschaden` erfolgt analog.

Im Anschluss soll untersucht werden, in wieweit Änderungen an der NLP-Pipeline Auswirkungen auf die Maße haben. Dazu bieten sich verschiedene Optionen an:

Auf Tokenebene

- Reduzierung der Wörter auf einen Wortstamm
- Reduzierung der Wörter auf ihre Wortart

[6] http://jgrapht.sourceforge.net
[7] von *recall*
[8] von *precision*

- Vereinheitlichung der Groß-/Kleinschreibung

Auf Dependenzgraphebene, unter anderem aus den Möglichkeiten des Stanford-Konverters resultierend

- Zusammenfassen von Multi-Token-Entitäten
- Ignorieren der Relationsrichtung
- Ignorieren der Relationsart (= Kantenbeschriftung)
- Nutzen der Modi „collapsing“ und „processing of conjunct dependencies“

Abschließend sollen typische Fehler analysiert und mögliche Verbesserungen diskutiert werden.

3.4 Anbindung an Equator

Die bei der Evaluation als beste abgeschnittene NLP-Pipeline soll in **Equator** eingebunden werden. Dazu wird die Editoransicht [Döh08, Abb. 3.8] um eine manuell aufrufbare Funktion erweitert, welche alle erkannten Entitäten markiert und die gefundenen Relationen anzeigt.

4 Verwandte Arbeiten

In [KYA07] wird ein umfassendes Informationssystem vorgestellt, welches einen ähnlichen Ansatz verfolgt. Aufgabe ist die fortlaufende automatische Erfassung, Auswertung und Darstellung von Nachrichten bezüglich eines Ereignisses wie einer Naturkatastrophe oder der Ausbruch einer Krankheit. Dazu werden in der NLP-Pipeline auch temporale und spatiale Informationen erfasst. Dies ermöglicht z.B. die räumliche und zeitliche Darstellung des Verlaufs einer Krankheitsausbreitung.

Die NER-Komponente des Systems erreicht ein P, R, $F1$-Maß von 87.60%, 87.80%, 87.70%. Die RE-Komponente kommt[9] bei der Klassifikation der binären Relationen (siehe 3.2.2) auf ein P, R, $F1$-Maß von 58.59%, 32.68%, 41.96%. Diese Zahlen deuten an, dass gerade im Bereich der Relationsextraktion noch Verbesserungspotenzial besteht.

Methodisch den gleichen Ansatz – Relationsextraktion über Muster in Dependenzgraphen – findet man in [Pie09]. Ziel ist hier die Identifizierung von Protein-Protein-Interaktionen

[9] Werte stammen vom Klassifikator mit dem höchsten $F1$-Maß.

(PPI) in biomedizinischen Texten. Die erreichten Werte[9] für die trinäre[10] PPI-Relation liegen bei einem *P*, *R*, *F1*-Maß von 43.74%, 48.77%, 46.12%.

Neben dem geplanten Verfahren, einen expliziten, generierten Musterkatalog für die Relationsextraktion zu verwenden, besteht auch die Möglichkeit, Kernel-Maschinen [RN04, S. 910ff] basierend auf Dependenzinformationen einzusetzen [RKP09].

[10]Der Autor untersucht zwar anfangs nur die binäre PP-Relation, durch weitere Maßnahmen (Interaktionswort-Filter) kommt aber das I hinzu.

Literatur

[BBC10] BBC News: *Haiti devastated by massive earthquake*. http://news.bbc.co.uk/2/hi/americas/8455629.stm. Version: Januar 2010. – Stand: 10.02.2010

[BM05] Bunescu, R.C. ; Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Morristown, NJ, USA : Association for Computational Linguistics, 2005, S. 724–731

[CEE+04] Carstensen, K.-U. ; Ebert, Ch. ; Endriss, C. ; Jekat, S. ; Klabunde, R. ; Langer, H.: *Computerlinguistik und Sprachtechnologie. Eine Einführung*. 2., überarb. u. erw. A. Spektrum Akademischer Verlag, 2004. – ISBN 9783827414076

[CJ05] Charniak, E. ; Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, 2005, S. 173–180

[CS07] Clegg, A.B. ; Shepherd, A.J.: Benchmarking natural-language parsers for biological applications using dependency graphs. In: *BMC Bioinformatics* 8 (2007), Nr. 1, S. 24+. http://dx.doi.org/10.1186/1471-2105-8-24. – DOI 10.1186/1471–2105–8–24

[Döh08] Döhling, L.: *Equator - Ein Wiki für die Task Force Erdbeben*. http://informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/studienDiplomArbeiten/finished/2008/doehling_studienarbeit_final.pdf. Version: Dezember 2008. – Studienarbeit

[FKZ07] Fundel, K. ; Küffner, R. ; Zimmer, R.: RelEx - Relation extraction using dependency parse trees. In: *Bioinformatics* 23 (2007), Nr. 3, S. 365–371. http://dx.doi.org/10.1093/bioinformatics/btl616. – DOI 10.1093/bioinformatics/btl616

[KYA07] Kawtrakul, A. ; Yingsaeree, C. ; Andrès, F.: A Framework of NLP Based Information Tracking and Related Knowledge Organizing with Topic Maps. In: *NLDB*, 2007, S. 272–283

[LZ06] Lemnitzer, L. ; Zinsmeister, H.: *Korpuslinguistik: Eine Einführung*. 1. Gunter Narr Verlag Tübingen, 2006. – ISBN 9783823362104

[Mit05] Mitkov, R. (Hrsg.): *The Oxford Handbook of Computational Linguistics*. Oxford University Press, USA, 2005 http://books.google.de/books?id=0aClhre-vW4C. – ISBN 9780199276349

[MMM06] Marneffe, M. ; MacCartney, B. ; Manning, C.D.: Generating Typed Dependency Parses from Phrase Structure Parses. In: *Proceedings of LREC-06*, 2006, S. 449–454

[MPK+05] McDonald, R. ; Pereira, F. ; Kulick, S. ; Winters, S. ; Jin, Y. ; White, P.: 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

[MS00] Manning, C.D. ; Schütze, H.: *Foundations of Statistical Natural Language Processing*. 2nd printing w. corrections. The MIT Press, 2000 http://nlp.stanford.edu/fsnlp. – ISBN 9780262133609

[Pie09] Pietschmann, S.: *Relationsextraktion durch Frequent Pattern in Dependency Graphen*. September 2009. – Diplomarbeit

[RKP09] Reichartz, F. ; Korte, H. ; Paass, G.: Dependency Tree Kernels for Relation Extraction from Natural Language Text. In: *ECML PKDD '09: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg : Springer-Verlag, 2009. – ISBN 978–3–642–04173–0, S. 270–285

[RN04] Russell, S. ; Norvig, P.: *Künstliche Intelligenz: Ein moderner Ansatz*. 2. Auflage. Pearson Studium, 2004

[U.S10a] U.S. Geological Survey: *Earthquakes with 1,000 or More Deaths since 1900*. http://earthquake.usgs.gov/earthquakes/world/world_deaths.php. Version: Februar 2010. – Stand: 02.02.2010

[U.S10b] U.S. Geological Survey: *Largest and Deadliest Earthquakes by Year*. http://earthquake.usgs.gov/earthquakes/eqarchives/year/byyear.php. Version: Februar 2010. – Stand: 02.02.2010

[U.S10c] U.S. Geological Survey: *Magnitude 7.0 - HAITI REGION - Earthquake Summary*. http://earthquake.usgs.gov/earthquakes/recenteqsww/Quakes/us2010rja6.php#summary. Version: Februar 2010. – Stand: 10.02.2010