Read the Web

TECHNISCHE UNIVERSITÄT DARMSTADT

Seminar aus maschinellem Lernen

Organisation: Frederik Janssen, Heiko Paulheim

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

- NELL – das Gesamtprojekt
- “Populating the Semantic Web by Macro-Reading Internet Text”
  - NLP-Ansatz für das Extrahieren von Informationen
  - Machbarkeit
- “Coupled Semi-Supervised Learning for Information Extraction”
  - Finden von Klassen und Relationen
- Projektausblick
- Fazit

TECHNISCHE UNIVERSITÄT DARMSTADT

NELL – DAS GESAMTPROJEKT

Projektkontext - Überblick

- Forschungsprojekt an der Carnegie Mellon University
- „Never-Ending Language Learner“
- Läuft seit Januar 2010 permanent

- Parsen von Webseiten mit unstrukturierten Informationen

- Ziele
  - Gewinnung von strukturierten Informationen
  - Datenbank mit gesammeltem Web-Wissen
  - Training und Anpassung der Vorgehensweise

Projektkontext - Details

TECHNISCHE UNIVERSITÄT DARMSTADT

- Initialisierung
  - Viele Kategorien
  - Viele Relationen
  - Jeweils 10 – 15 Instanzen

- Datensatz: „ClueWeb09“
  - Nicht komprimiert 25TB
  - Über 1 Mrd. Webseiten
  - 10 Sprachen
  - Links untereinander und nach extern

TECHNISCHE UNIVERSITÄT DARMSTADT

POPULATING THE SEMANTIC WEB BY MACRO-READING INTERNET TEXT

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zukunft des Semantik Web hängt von dessen Inhalt ab
  - Umfang
  - Detailtreue

- Wie kommt man zu diesem Inhalt?
  - Manuelles Hinzufügen von Informationen
  - Anpassung bereits existierender Datenbanken
  - Software, die automatisch unstrukturierten Text parst

Natural Language Processing

TECHNISCHE UNIVERSITÄT DARMSTADT

- NLP-Methoden State-of-the-art

- Komplett verstehen, was ein Text aussagt, kann noch kein Algorithmus

- Ist im Internet aber nicht unbedingt nötig
  - Macro-reading
  - Ontology-driven reading
  - Machine learning methods

Macro-reading

TECHNISCHE UNIVERSITÄT DARMSTADT

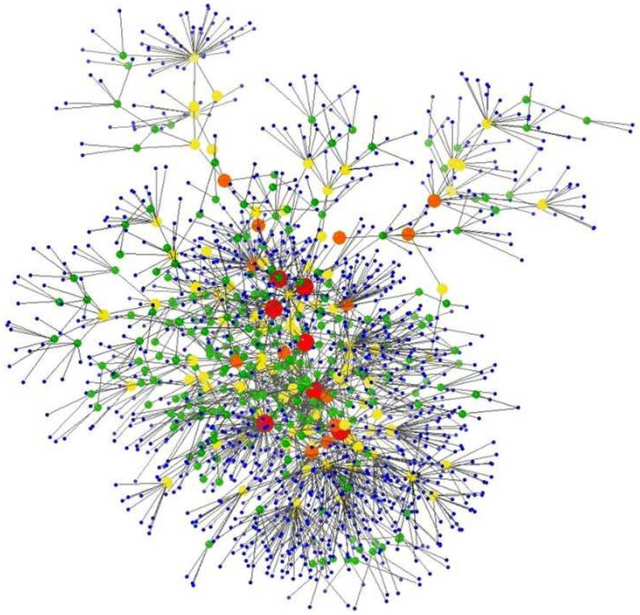
	Micro-reading	**Macro-reading**
Eingabe:	*1 Text*	*Viele Texte (z.B. das Web)*
Ausgabe:	*Komplettes daraus extrahierbares Wissen*	*Einige extrahierbare Fakten*

- Vorteile des Macro-reading
  - Es wird nicht erwartet, dass alle Informationen erfasst werden
  - Redundanz erleichtert das Extrahieren

Ontology-driven reading

TECHNISCHE UNIVERSITÄT DARMSTADT

- Nur relevante Daten werden geparst
  - Im Bezug auf die gegebene Ontologie
  - und die bisher darin gefundenen Instanzen



Machine learning methods

TECHNISCHE UNIVERSITÄT DARMSTADT

- Finden von Mustern
  - „mayor of X“ → X ist eine Stadt
  - Für jede initiale Kategorie und Relation
  - Weitere Spezialisierung durch Nebenbedingungen
    - Type Checking:
    - A plays-on-team B
    - Nur wenn A eine Person
    - und B ein Team ist
  - Komplexer → genauer

HTML pattern learner

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zusätzlich interpretieren von HTML-Strukturen
  - Z.B. Liste → Aufzählung ähnlicher Instanzen

- Beispiel: Fußballmannschaften
  - FC Bayern
  - VfB Stuttgart
  - …

Ende einer Iteration

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aufnahme neuer Fakten
  - Nur wenn Text-Methode
  - und Struktur-Methode matchen

Skype
isA: company
company_economic_sector: VoIP
competes_with: AOL, MSN, Yahoo, Google
acquired_by: Ebay

EBay
isA: company
company_CEO: Pierre Omidyar
competes_with: Dell, Google, Yahoo,
Amazon, Amazon.com, Microsoft, AOL
acquired: PayPal, Skype

- Nächste Iteration
  - endlos

Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

- Experiment 1
  - 16 initiale Kategorien
  - 4224 Instanzen von Kategorien gefunden (97% Precision)

- Experiment 2
  - 16 initiale Kategorien
  - 14 initiale Relationen
  - 15520 Instanzen von Kategorien gefunden (83% Precision)
  - 2674 Instanzen von Relationen gefunden (84% Precision)

Ausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Machbarkeit gezeigt

- Weitere Forschung
  - Weitere Optimierung
  - Micro-reader mittels Macro-reader trainieren
    - Weil diese wichtig für seltenere Daten sind (z.B. Personendaten)
  - Macro-reader auch für weniger redundante Daten optimieren
  - Semi-supervised Learning

Kritik

TECHNISCHE UNIVERSITÄT DARMSTADT

- Seiten in anderen Sprachen
  - Werden verwendet
  - Aber inwiefern geparst? → Ergebnisse?
  - Und inwiefern auf dieselbe Weise umsetzbar?

- HTML pattern learner
  - Keine weiteren Beispiele, außer Listen

- Precision
  - Verbesserungsbedürftig

TECHNISCHE UNIVERSITÄT DARMSTADT

COUPLED SEMI-SUPERVISED LEARNING FOR INFORMATION EXTRACTION

Semi-supervised learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- Wenige statt vielen Trainingsbeispielen
  - Supervised learning ist teuer
- Große Menge nicht gelabelter Text

- Problem
  - Fehler akkumulieren sich
  - Mäßige Precision

- Lösungsansatz
  - Kopplung mit Nebenbedingungen

Kopplung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gegenseitiges Ausschließen
  - Bestimmte Prädikate schließen sich gegenseitig aus
  - Gegeben bei Initialisierung
- Type Checking
  - A plays-on-team B
  - Nur wenn A eine Person und B ein Team ist
- Unabhängiges Matching
  - der NLP- und der HTML tag-Varianten

Coupled Pattern Learner (CPL)

TECHNISCHE UNIVERSITÄT DARMSTADT

Input: An ontology *O*, and text corpus *C*
Output: Trusted instances/contextual patterns for each predicate

for *i* = 1, 2, …, ∞ **do**
  foreach *predicate p* ∈ *O* **do**
    EXTRACT new candidate instances/contextual patterns using
      recently promoted patterns/instances;
    FILTER candidates that violate coupling;
    RANK candidate instances/patterns;
    PROMOTE top candidates;
  end
end

CPL – Einzelschritte

TECHNISCHE UNIVERSITÄT DARMSTADT

- EXTRACT
  - Finden von neuen Instanzen
    - Mit Hilfe der Pattern aus der letzten Iteration
    - Auswahl der 1000 häufigsten passenden Instanzen
  - Finden von neuen Pattern
    - Mit Hilfe von Instanzen aus der letzten Iteration
- FILTER
  - Instanz wird nur aufgenommen, wenn sie mindestens 3x häufiger mit Pattern A als mit Pattern B gefunden wird (Pattern A und Pattern B schließen sich gegenseitig aus)
  - Bei Finden von Pattern ähnlich

CPL – Einzelschritte (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

- RANK
  - Ranking gefundener Instanzen/Pattern nach Precision
  - $Precision(p) = \frac{\sum i \epsilon I \; count(i,p)}{count(p)}$
  - $count(i,p)$ Anzahl, wie oft die Instanz i mit dem Pattern p vorkommt
  - $count(p)$ Anzahl, wie oft das Pattern p insgesamt vorkommt
- PROMOTE
  - Hinzunehmen bester Instanzen und Pattern
  - Maximal 100 Instanzen und 5 Pattern
  - Instanz muss mindestens in 2 Pattern vorkommen

Coupled SEAL (CSEAL)

TECHNISCHE UNIVERSITÄT DARMSTADT

Input: An ontology *O*, and text corpus *C*
Output: Trusted instances/wrappers for each predicate

for *i* = 1, 2, …, ∞ **do**
  foreach *predicate p* ∈ *O* **do**
    begin Call existing SEAL code to:
      QUERY for documents containing recently promoted instances;
      LEARN wrappers for each document returned;
      EXTRACT new candidates using wrappers;
    end
    FILTER wrappers that extract candidates that violate coupling;
    RANK candidate instances;
    PROMOTE top candidates;
  end
end

CSEAL – Einzelschritte

TECHNISCHE UNIVERSITÄT DARMSTADT

- SEAL findet Instanzen in HTML-Strukturen
  - Bietet aber keine Möglichkeit, Kopplung zu integrieren
- FILTER
  - Entfernen aller Dokumente, die Instanzen sich ausschließender Prädikate finden
  - Type Checking
- RANK
  - Ranking wie bei CPL
- PROMOTE
  - Hinzunehmen wie bei CPL
  - Pro Domain wird maximal eine Seite verwendet

Meta-Bootstrap Learner (MBL)

TECHNISCHE UNIVERSITÄT DARMSTADT

Input: An ontology *O*, a set of extractors *E*
Output: Trusted instances for each predicate

for *i* = 1, 2, …, ∞ **do**
  foreach *predicate p* ∈ *O* **do**
    foreach *extractor e* ∈ *E* **do**
      EXTRACT new candidates for *p* using *e* with recently promoted instances;
    end
    FILTER candidates that violate mutual-exclusion or type-checking constraints;
    PROMOTE candidates that were extracted by all extractors;
  end
end

Experiment - Methodik

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontologie
  - Rund um *companies* und *sports*
    - und ein paar andere (Negativbeispiele)
  - Jeweils 15 Instanzen und 5 Pattern pro Kategorie
  - Jeweils 15 Instanzen und 5 Negativ-Instanzen pro Relation
  - Meiste Prädikate definiert sich gegenseitig ausschließend
- 200 Mio. Seiten wurden gefiltert (englisch, jugendfrei)
- 10 Iterationen

Predicate	Precision (%)					Promoted Instances (#)				
	CPL	UPL	CSEAL	SEAL	MBL	CPL	UPL	CSEAL	SEAL	MBL
AcademicField	70	83	90	97	100	46	903	203	1000	181
Actor	100	33	100	97	100	199	1000	1000	1000	380
Animal	80	50	90	70	97	741	1000	144	974	307
Athlete	87	17	100	87	100	132	930	276	1000	555
AwardTrophyTournament	57	7	53	7	77	86	902	146	1000	79
BoardGame	80	13	70	77	90	10	907	126	1000	31
BodyPart	77	17	97	63	93	176	922	80	1000	61
Building	33	50	30	0	93	597	1000	57	1000	14
Celebrity	100	90	100	100	97	347	1000	72	747	514
CEO	33	30	100	77	100	3	902	322	1000	30
City	97	100	97	87	97	1000	1000	368	1000	603
Clothing	97	20	43	27	97	83	973	167	1000	102
Coach	93	63	100	83	100	188	838	619	1000	242
Company	97	83	100	100	97	1000	1000	245	1000	784
Conference	93	53	97	90	100	95	990	437	928	92
Country	57	33	97	37	93	1000	1000	130	1000	207
EconomicSector	60	23	100	10	77	1000	1000	34	1000	138
Emotion	77	53	87	60	83	483	992	183	1000	211
Food	90	70	97	80	100	811	1000	89	1000	272
Furniture	100	0	57	57	90	55	963	215	1000	95
Hobby	77	33	77	50	90	357	936	77	1000	127
KitchenItem	73	3	88	13	100	11	900	8	960	2
Mammal	83	50	93	50	90	224	1000	154	1000	169
Movie	97	57	97	100	100	718	1000	566	1000	183
NewspaperCompany	90	60	60	97	100	179	1000	1000	1000	241
Politician	80	60	97	37	100	178	990	30	1000	101
Product	90	83	-	77	70	1000	1000	0	999	127
ProductType	73	63	27	63	50	712	1000	31	1000	159
Profession	73	53	-	57	93	916	973	0	1000	171
ProfessionalOrganization	93	63	100	77	87	104	943	58	1000	163
Reptile	95	3	90	27	100	19	912	149	1000	54
Room	64	0	33	7	100	25	913	12	643	3
Scientist	97	30	100	17	100	83	971	928	1000	130
Shape	77	7	7	7	85	43	985	28	733	26
Sport	77	13	63	83	73	283	1000	225	1000	284
SportsEquipment	20	10	57	23	23	58	902	52	1000	174
SportsLeague	100	7	80	27	86	11	901	10	1000	14
SportsTeam	90	30	87	87	87	301	903	864	944	506
Stadium	93	57	53	63	90	102	767	944	1000	343
StateOrProvince	77	63	83	93	77	202	1000	114	1000	161
Tool	40	13	93	90	97	561	1000	713	1000	59
Trait	53	40	52	47	97	234	1000	21	1000	44
University	93	97	100	90	93	1000	1000	961	1000	516
Vehicle	67	30	50	13	77	460	1000	50	1000	98
Average	78	41	78	59	90	360	960	271	976	199
Weighted average	79	42	86	59	91					

Experiment - Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

	Precision (%)					Promoted Instances (#)				
Predicate	CPL	UPL	CSEAL	SEAL	MBL	CPL	UPL	CSEAL	SEAL	MBL
CompanyAcquiredCompany	97	77	-	-	-	93	230	0	0	0
AthletePlaysForTeam	100	93	100	76	100	9	269	4	17	96
AthletePlaysInLeague	-	78	100	57	-	0	18	14	82	0
AthletePlaysSport	100	47	100	100	100	83	258	1	1	109
CEOOfCompany	100	100	-	100	100	18	18	0	1	1
CityLocatedInCountry	93	57	100	100	100	185	787	9	577	136
CityLocatedInState	100	70	100	93	100	76	194	34	537	54
CoachCoachesInLeague	-	-	0	-	-	0	0	1	0	0
CoachCoachesTeam	100	100	-	-	100	324	668	0	0	6
CompanyIsInEconomicSector	93	97	-	-	-	583	889	0	0	0
CompanyCompetesWithCompany	100	67	-	-	-	28	123	0	0	0
CompanyHasOfficeInCity	-	63	-	100	-	0	526	0	4	0
CompanyHasOfficeInCountry	-	90	-	-	-	0	195	0	0	0
CompanyHeadquarteredInCity	50	53	100	100	-	2	532	1	2	0
LeaguePlaysGamesInStadium	-	-	-	100	-	0	0	0	177	0
CompanyProducesProduct	97	93	-	-	100	54	215	0	0	8
ProductInstanceOfProductType	73	67	-	-	-	153	484	0	0	0
SportUsesSportsEquipment	33	3	100	87	33	15	1330	5	15	6
StadiumLocatedInCity	100	20	77	70	90	7	600	200	554	56
StateHasCapitalCity	60	70	-	73	-	266	188	0	495	0
StateLocatedInCountry	97	40	100	97	100	194	1299	46	653	61
TeamHasHomeStadium	100	87	100	100	100	97	208	179	106	92
TeamPlaysAgainstTeam	100	80	-	-	-	238	2088	0	0	0
TeamHasHomeCity	-	57	-	93	100	0	680	0	29	11
TeamPlaysInLeague	100	67	100	100	100	7	255	104	749	23
TeamPlaysSport	-	70	100	100	100	0	177	30	30	37
TeamWonAwardTrophyTournament	90	70	-	-	-	128	262	0	0	0
Average	89	69	91	91	95	95	463	23	149	26
Weighted Average	91	61	92	90	99					

Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

- CPL schneidet besser ab als UPL
- CSEAL schneidet besser ab als SEAL
- MBL schneidet besser ab als CPL oder CSEAL alleine
- Das Einbeziehen der Kopplung lohnt sich also

Ausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Weitere Forschung
  - Synonym-Auflösung bzw. Bedeutungsanalyse (NLP)
  - Wann soll der Algorithmus terminieren?
  - Bei zu allgemeinen Begriffen müssen Probleme beseitigt werden
    - Ontologie eventuell ausbauen?
  - Recall verbessern
  - Weitere Einschränkungen könnten helfen
    - Z.B. das Wissen, dass ein Land nur eine Hauptstadt hat
  - Für manche Relationen wurden keine Instanzen gefunden
    - Verbesserung durch Lernalgorithmen möglich?

Kritik

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gut: Sehr ausführliche Beschreibung der Methodik
- Wie lange dauert eine Iteration?
- Viele Wiederholungen in einzelnen Abschnitten
  - Überflüssig, wenn man Paper am Stück liest
- Kombination von UPL/SEAL oder z.B. CPL/SEAL
  - Wie würden die Ergebnisse hier aussehen?

TECHNISCHE UNIVERSITÄT DARMSTADT

PROJEKTAUSBLICK

Projektausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- In der Zwischenzeit wurden 8 weitere Paper veröffentlicht
  - Auflösen von Synonymen
  - Finden von Abhängigkeiten zwischen Kategorie-Instanzen
  - Veränderung von Wortbedeutungen über die Jahre
  - Weitere Verbesserungen
- Es wird weiter intensiv geforscht
  - Neuestes Paper von 2012
- NELL ist etwa in der 500. Iteration

TECHNISCHE UNIVERSITÄT DARMSTADT

FAZIT

Fazit

TECHNISCHE UNIVERSITÄT DARMSTADT

- Reihenfolge unklar
  - Erstes Paper zitiert zweites Paper
- Ansatz für Daten wie das Internet sinnvoll
  - Allgemeinwissen
  - Weit verbreitete Fakten
- Semantik Web
  - Dieses Problem teilweise gelöst
  - Rückt näher

TECHNISCHE UNIVERSITÄT DARMSTADT

Vielen Dank für die Aufmerksamkeit!

FRAGEN?

Quellen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Texte
  - Populating the Semantic Web by Macro-Reading Internet Text (2009; T.M. Mitchell, J.Betteridge, A. Carlson, E.R. Hruschka Jr. and R.C. Wang)
  - Coupled Semi-Supervised Learning for Information Extraction (2010; A. Carlson, J. Betteridge, R.C. Wang, E.R. Hruschka Jr. and T.M. Mitchell)
- Bilder
  - http://www.cmu.edu/homeimages/CarnegieMellonUniversity_wordmark.gif (Folie 4)
  - http://lemurproject.org/logo/lemur_logo_50.gif (Folie 5)
  - http://www.w3c.it/talks/2009/athena/images/sw-vert-w3c.jpg (Folie 7)
  - http://lsdis.cs.uga.edu/projects/glycomics/report/webpage/graphics/ontology.jpg (Folie 10)