Proseminar Linguistische Annotationen

Semi-automatische Annotation

Patricia Helmich

SS 2010

Gliederung

- Manuelle vs. semi-automatische Annotation
- Penn Chinese Treebank
  - Nutzen von automatischen Tools bei der Annotation
  - Nutzen von automatischen Tools bei der Annotationsqualitätskontrolle
- BioProp
  - Nutzen eines domänenfremden *Semantic Role Labelers*
  - Entwicklung eines automatischen *Semantic Role Labelers für die biomedizinische Domäne*

Manuelle Annotation

- Ein (menschlicher) Annotator bekommt ein nicht bearbeitetes Korpus und annotiert dieses manuell und oft von Grund auf
- Hilfen: sein Basiswissen, linguistisches Training, Leitfaden

→ SEHR zeitaufwändig

Lösung: semi-automatische Annotation

Semi-automatische Annotion

- Korpus wird von einem automatischen System annotiert, welches vorher auf anderen Korpora trainiert wurde

- Ein menschlicher Annotator überarbeitet danach die Ergebnisse des Systems:

  Er überprüft, ob die Tags richtig gewählt wurden, und korrigiert diese, wenn nötig

- Die Annotationsaufgabe wird also in eine Korrekturaufgabe umgewandelt, wodurch viel Zeit gespart werden kann

Penn Chinese Treebank (CTB)

- Projekt mit dem Ziel, ein chinesisches Korpus aufzubauen, welcher vollständig segmentiert, mit POS-Tags annotiert und syntaktisch geklammert ist

- Die erste Version (CTB-I) besteht aus einem Korpus der Xinhua Nachrichtenagentur aus den Jahren 1994 bis 1998, sie umfasst 100 000 Wörter

- Zur Zeit (2002) Entwicklung der zweiten Version (CTB-II), durch Existenz von CBT-I neue Herangehensweise

- Auf Basis von CBT - I als Trainingsmaterial können neue automatische *chinese language processing (CLP)* Tools trainiert werden

  → diese werden benutzt, um die zu annotierenden Texte für die Entwicklung des CBT-II vorzubearbeiten

Annotationstempo

Das Annotationstempo wird hauptsächlich durch 3 Faktoren beeinflusst:

- Hintergrundwissen des Annotators
- Design des Leitfadens
- Verfügbarkeit von vorverarbeitenden Tools

Automatische Tools

- Durch die Verfügbarkeit von CTB-I ist es möglich ein Anzahl von CLP-Tools mit immer höherer Akkuratheit zu trainieren

- Bei Verwendung dieser Tools als vorverarbeitende Methode wird der Annotationsprozess der Korpora für CTB-II sehr beschleunigt

Segmentierung chinesischer Wörter

- Wortkomponenten (chinesische Zeichen) können am linken Rand, in der Mitte und am rechten Rand eines Wortes stehen

- Je nach Position können die Komponenten verschiedene Bedeutungen haben

  → Segmentierungsaufgabe kann als Disambiuierungsaufgabe betrachtet werden

- Training eines Automatischen Wortsegmentierers durch Verwendung der Daten von CTB-I

  → Taggen der Komponenten als LL / MM / RR / LR

- Trainingskorpus: 80 000 Wörter aus CTB-I

- Testkorpus: 20 000 Wörter aus CBT-I

- Segmentierer erzielt eine Akkuratheit von 91%

POS - Tagger

- Auf segementierte Sätze kann ein POS-Tagger angewendet werden, ähnlich wie für indoeuropäischen Sprachen
- Kontexte, welche POS-Tags voraussagen, sind im Chinesischen und Englischen in etwa gleich: Nachbarwörter, Informationen über vorherige Tags und die Wortkomponenten
- Unterschied zum Indoeuropäischen: die fehlende Präfix- und Suffixmorphologie im Chinesischen (guter Indikator für POS)
- Selbes Trainings- bzw. Testkorpus wie für den Segmentierer
- Akkuratheit des POS-Taggers: 93%

Automatische Tools

- Entwicklung und Training der chinesischen Segmentierer und Tagger beschleunigt Annotation

- Gleichzeitig ist es wiederum durch mehr annotierte Daten möglich, Tools mit noch höherer Akkuratheit zu trainieren

  → Bootstrapping cycle,
  sowohl für die Annotation als auch für die Tools nützlich

- Korrektur des Outputs eines Segmentierers oder eines POS-Taggers fast zweimal so schnell wie Annotation von Grund auf

Statistical Parser

Der Nutzen eines Parsers als vorverarbeitendes Tool ist weniger offensichtlich als der eines Segmentierers oder POS-Taggers

→ bei einem Fehler des Parsers muss der menschliche Annotator teilweise viel Backtracking betreiben

Frage: Lohnt sich der Einsatz eines automatischen Parsers überhaupt?

Automatische Tools

Statistical Parser

- Experiment zur Klärung der Frage:
  - Ein auf 80 000 Wörtern trainierter und auf 10 000 Wörtern aus CBT-I getesteter Parser erzielt 73,9% Precision und 72,2% Recall
  - Ein zufällig ausgewählter, ca. 13 500 Wörter umfassender Teil des Korpus wurde ausgewählt und in 2 Hälften geteilt
  - Die eine Hälfte wurde von einem Annotator von Grund auf annotiert, die andere von dem Parser vorverarbeitet und dann von einem Annotator korrigiert
  - Beide Teile werden von zweitem Annotator gecheckt
    → adjudizierte Daten Goldstandard

Statistical Parser

Ergebnisse:

Teil	Precision	Recall	Zeit/h	Akkuratheit
1	N/A	N/A	28:01	99,84%
2	76,73	75,36	16:21	99,76%

→ Reduzierung der Bearbeitungszeit um 42%

→ Einsatz des Parsers lohnt sich trotz Backtracking

Qualitätskontrolle der Annotation

- Evaluationstools helfen bei der Überwachung der Akkuratheit der Annotation und der Interannotatorkonsistenz, vor allem in der Phase der syntaktischen Klammerung

- Selbst die besten Annotatoren machen Fehler, vor allem mechanische Fehler

  → Gerade solche mechanischen Fehler sind mit automatischen Tools gut zu finden

- Kontrolle der Akkuratheit und der Interannotatorkonsistenz:
  - 20% der Daten werden ausgewählt und doppelt annotiert
  - Wöchentlicher Vergleich der Annotationen in drei Schritten:
    (1) Ein Evaluationtool läuft über jede doppelt annotierte Datei und ermittelt die Interannotatorkonsistenz
    (2) Die Annotatoren untersuchen die Ergebnisse und die Inkonsistenzen, die das System ermittelt hat, und korrigieren/adjudizieren diese (korrigierte/adjudizierte Daten → Goldstandard)
    (3) Vergleich des Goldstandards mit der Annotation eines jeden Annotators, um die Akkuratheit der Annotatoren zu bestimmen

  → Beide Ergebnisse liegen in der oberen Hälfte der 90%

Penn Chinese Treebank - Überblick

Einsatz von automatischen Tools:

- Segmentierung chinesischer Wörter (Akkuratheit von 91%)
- POS-Tagger (Akkuratheit von 93%)
- Statistischer Parser (Akkuratheit von 99,76%)

→ Zeitverringerung jeweils fast um die Hälfte

- Qualitätskontrolle der Annotation:

  Aufdecken von Inkonsistenzen der Annotatoren sowie mechanischen Annotationsfehlern

BioProp

- Eine biomedizinische Proposition Bank
- Enthält Annotationen zu Prädikat-Argument-Strukturen (PAS) und semantischen Rollen im Treebank-Schema (wie PropBank für die Nachrichtendomäne)
- Anwendung:
  - Fokus in Informationsextraktionssystemen verschiebt sich zur Zeit (2006) von Information über *Named Entities* (NE) zu Information über Relationen zwischen NE‘s
  - Bei Informationsextraktion (IE) von Relationen müssen die semantischen Rollen der NE‘s (*predicate arguments*) zum zugehörigen Verb (*predicate*) bestimmt werden:

    Semantic role labeling (SRL)
  - ein automatisches SRL-System soll entwickelt werden, welches auf einem biomedizinischen annotierten Korpus (Proposition Bank) trainiert werden soll

BioProp - Annotation

- Die manuelle Annotation der PAS‘s zur Entwicklung der Proposition Bank ist sehr zeitaufwändig
- Aufgrund der Komplexität der Proposition Bank: häufig auftretende Inkonsistenzen bei der Annotation
- Trotzdem: es gibt adäquate Proposition Banks in der Nachrichtendomäne zum Training von SRL-Systemen
- Die Leistung eines SRL-Systems verschlechtert sich nicht signifikant, wenn mit dem System domänenfremde Korpora getaggt werden

  (Carreras und Márquez, 2005)

→ Lösung: Annotation des biomedizinischen Korpus mit Hilfe eines SRL-Systems aus der Nachrichtendomäne (Wall Street Journal (WSJ) SRL-System) und dann Überarbeitung durch menschliche Annotatoren

BioProp - Annotation

- Biomedizinisches Korpus: GENIA, eine Sammlung von MEDLINE Abstracts
  - Penn-style Treebank: GENIA Treebank, annotiert mit POS-Tags
  - Im Gegensatz zum WSJ-Korpus: keine Proposition Bank für GENIA verfügbar

- Auswahl von 30 Verben nach Häufigkeit ihrer Vorkommen und Bedeutsamkeit in biomedizinischen Texten

- Fokus liegt auf Verwendung der Verben im biomedizinischen Kontext

Framesets der biomedizinischen Wörter

- Basiert auf den Framesets von VerbNet, teilweise angepasst
- Teilweise Unterschiede in der Nutzung der Verben im allgemeinen Kontext vs. biomedizinischen Kontext

→ Unterteilung der Verben in vier Typen:

(1) Verben, die nicht in VerbNet enthalten sind, da sie in der Nachrichtendomäne selten auftreten
(2) Verben, die in VerbNet enthalten sind, deren biomedizinische Bedeutungen und Framesets jedoch in VerbNet nicht definiert sind
(3) Verben, die in VerbNet enthalten sind, inklusive biomedizinischer Bedeutungen, deren Hauptbedeutung jedoch unterschiedlich definiert für die beiden Domänen
(4) Verben mit gleicher Verwendung in beiden Domänen

Framesets der biomedizinischen Wörter

- Verben vom Typ (1):
  - Nach der Analyse aller Sätze, welche ein Verb dieses Typs enthalten, wird das Frameset für dieses Verb definiert, z. B. für phosphorylate
  - Manchmal können Framesets von anderen Verben in VerbNet übernommen werden, z. B. „transactivate“ kann Frameset von „activate“ übernehmen
- Verben vom Typ (2):
  - Framesets für die zusätzlichen biomedizinischen Bedeutungen müssen hinzugefügt werden,
    z. B. ist „express“ in VerbNet als „say“ bzw. „send very quickly“ definiert,
    im biomedizinischen Kontext bedeutet es dagegen eher „translate“
  - Hier können häufig Framesets von anderen Verben in VerbNet übernommen werden

Framesets der biomedizinischen Wörter

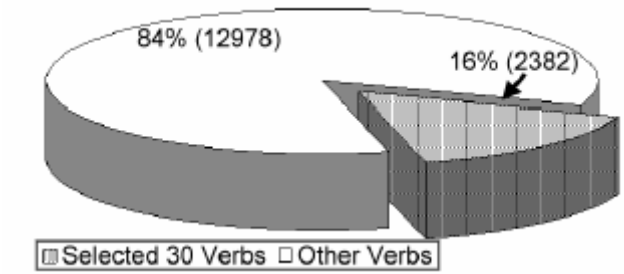
- Verben vom Typ (3):
  - Für diese Verben müssen nur die schon vorhandenen Framesets als Hauptbedeutung eingesetzt werden,

    Bsp.: „bind“, was in Nachrichtentexten eher im Sinne von „to tie“ benutzt wird, in biomedizinischen Texten dagegen in der Bedeutung „to attach“

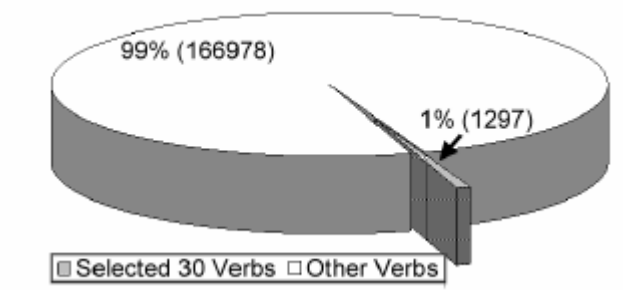
- Verben vom Typ (4):
  - Hier können die Framesets aus der Nachrichtendomäne direkt übernommen werden, da diese identisch sind

Verteilungen der 30 ausgewählten Verben

Die Verteilung der 30 Verben und allen anderen Verben in BioProp:



Die Verteilung der 30 Verben und allen anderen Verben in PropBank:



Verteilungen der 30 ausgewählten Verben

Type	Verb list
1	encode, interact, phosphorylate, transactivate
2	express, modulate
3	bind
4	activate, affect, alter, associate, block, decrease differentiate, encode, enhance, increase, induce, inhibit, mediate, mutate, prevent, promote, reduce, regulate, repress, signal, stimulate, suppress, transform, trigger

Verbs	# in BioProp	Ratio(%)	# in PropBank	Ratio(%)
induce	290	1.89	16	0.01
bind	252	1.64	0	0
activate	235	1.53	2	0
express	194	1.26	53	0.03
inhibit	184	1.20	6	0
increase	166	1.08	396	0.24
regulate	122	0.79	23	0.01
mediate	104	0.68	1	0
stimulate	93	0.61	11	0.01
associate	82	0.53	51	0.03
encode	79	0.51	0	0
affect	60	0.39	119	0.07
enhance	60	0.39	28	0.02
block	58	0.38	71	0.04
reduce	55	0.36	241	0.14
decrease	54	0.35	16	0.01
suppress	38	0.25	4	0
interact	36	0.23	0	0
alter	27	0.18	17	0.01
transactivate	24	0.16	0	0
modulate	22	0.14	1	0
phosphorylate	21	0.14	0	0
transform	21	0.14	22	0.01
differentiate	21	0.14	2	0
repress	17	0.11	1	0
prevent	15	0.10	92	0.05
promote	14	0.09	52	0.03
trigger	14	0.09	40	0.02
mutate	14	0.09	1	0
signal	10	0.07	31	0.02

BioProp - Annotation

Annotationsprozess:

(1) Identifizierung der Prädikatkandidaten

(2) Automatische Annotation der biomedizinischen semantischen Rollen mit dem WSJ SRL-System

(3) Transformation der automatischen Tagging-Ergebnisse ins WordFreak-Format

(4) Manuelle Korrektur der Annotationsergebnisse mit dem WordFreak-Tool

Inter-Annotation Agreement

- Konsistenztest auf 2382 Instanzen biomedizinischer Propositionen

Kappa-Statistik: $\kappa = \frac{A_o - A_e}{1 - A_e}$

		$P(A)$	$P(E)$	Kappa score
including ArgM	role identification	.97	.52	.94
	role classification	.96	.18	.95
	combined decision	.96	.18	.95
excluding ArgM	role identification	.97	.26	.94
	role classification	.99	.28	.98
	combined decision	.99	.28	.98

Arbeitsaufwand der Annotation

- Arbeitseinsparung durch Nutzen des WSJ SRL-System, welches dies semantischen Rollen automatisch annotiert

  → menschlicher Annotator muss nur bei den meisten Tags nur entscheiden, ob sie richtig oder falsch sind und nicht alle Wörter selbst taggen

- Maß für den Grad der Arbeitseinsparung:

$$\rho = \frac{\#\text{ of correctly labeled nodes}}{\#\text{ of all nodes}} < \frac{\#\text{ of correctly labeled nodes}}{\#\text{ of correct} + \#\text{ of incorrect} + \#\text{ of missed nodes}}$$

- Arbeitseinsparung für BioProp: $\rho < \frac{18932}{18932 + 6682 + 15316} = \frac{18932}{40975} = 46\%$

Effekt der biomedizinischen Trainingskorpora auf SRL-Systems

- Konstruktion eines neues SRL-Systems auf Basis von BioProp:

  BIOSMILE (BIOmedical SeMantIc roLe labEler)

- SRL in zwei Schritten:
  - Alle Prädikate identifizieren
  - Für jedes Prädikat alle zugehörigen Argumente markieren
- Features, welche für das Argumentklassifizierungsmodell genutzt wurden:

BASIC FEATURES

- **Predicate** – The predicate lemma
- **Path** – The syntactic path through the parsing tree from the parse constituent being classified to the predicate
- **Constituent type**
- **Position** – Whether the phrase is located before or after the predicate
- **Voice** – passive: If the predicate has a POS tag VBN, and its chunk is not a VP, or it is preceded by a form of "to be" or "to get" within its chunk; otherwise, it is active
- **Head word** – Calculated using the head word table described by Collins (1999)
- **Head POS** – The POS of the Head Word
- **Sub-categorization** – The phrase structure rule that expands the predicate's parent node in the parsing tree
- **First and last Word and their POS tags**
- **Level** – The level in the parsing tree

PREDICATE FEATURES

- **Predicate's verb class**
- **Predicate POS tag**
- **Predicate frequency**
- **Predicate's context POS**
- **Number of predicates**

FULL PARSING FEATURES

- **Parent's, left sibling's, and right sibling's paths, constituent types, positions, head words and head POS tags**
- **Head of PP parent** – If the parent is a PP, then the head of this PP is also used as a feature

COMBINATION FEATURES

- **Predicate distance combination**
- **Predicate phrase type combination**
- **Head word and predicate combination**
- **Voice position combination**

OTHERS

- **Syntactic frame of predicate/NP**
- **Headword suffixes of lengths 2, 3, and 4**
- **Number of words in the phrase**
- **Context words & POS tags**

Effekt der biomedizinischen Trainingskorpora auf SRL-Systems

- Experiment zur Überprüfung, ob Training des SRL-Systems auf biomedizinischem Korpus vs. Korpus mit Zeitungstexten (Domäne: Finanzen/Wirtschaft) einen bedeutenden Effekt hat:
  - BIOSMILE wird auf 30 zufällig ausgewählten Trainingssets aus BioProp (g1,…,g30) trainiert
  - WSJ SRL-System wird auf 30 Trainingssets von PropBank trainiert (w1,…,w30)
  - Beide Systeme werden auf 30 Testsets von BioProp getestet (g1 und w1 auf Testset 1, etc.)

Training	Test	Precision	Recall	F-score
PropBank	BioProp	74.78	56.25	64.20
BioProp	BioProp	88.65	85.61	87.10

  - BIOSMILE übertrifft WSJ SRL-System um 22,9%, ein statistisch signifikantes Ergebnis

Überblick BioProp

- Domänenfremdes SRL-System erzielt gute Ergebnisse bei der Annotation des biomedizinischen Korpus
- Inter-Annotator Agreement: Kappa-Score = 0,95
- Zeiteinsparung um höchstens 46%
- Entwicklung eines SRL-Systems für die biomedizinische Domäne verbessert den F-Score für die Annotation eines biomedizinischen Textes um 22,9% im Vergleich zum domänenfremden SRL-System
- **Aber:** Die meisten Verben haben nur eine Bedeutung!
  - → Erleichtert Annotation und erklärt auch hohes Inter-Annotator Agreement
  - → Frage: Ebenso gute Ergebnisse bei mehrdeutigen Verben?

Literaturangaben:

- Chou, W., Tsai, R. T., Su, Y., Ku, W., Sung, T., and Hsu, W. 2006. A semi-automatic method for annotating a biomedical proposition bank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006* (Sydney, Australia, July 22 - 22, 2006). ACL Workshops. Association for Computational Linguistics, Morristown, NJ, 5-12.

- Xue, N., Chiou, F., and Palmer, M. 2002. Building a large-scale annotated Chinese corpus. In *Proceedings of the 19th international Conference on Computational Linguistics - Volume 1* (Taipei, Taiwan, August 24 - September 01, 2002). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-8.