Proseminar Computerlinguistik

Wintersemester 2016/17

Ahmed Albawabiji

Kurzes Referat über das Thema

Inter-Annotator Agreement for a German Newspaper Corpus, Thorsten Brants, Jahr 2000

Studiengang Linguistische Informatik
ahmed.albawabiji@fau.de

Erlangen, 02.02.2016

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Inhaltverzeichnis:

Einführung: Worum geht es?

- Untersuchung an einem NEGRA Korpus, das aus Texten deutscher Zeitungen besteht.
- Dieses Korpus ist syntaktisch mit sowohl POS als auch struktureller Information annotiert.
- Agreement(Zustimmung) für POS ist 98.6% , labeled F- Score for Structures 92.4%

Figur:(1)

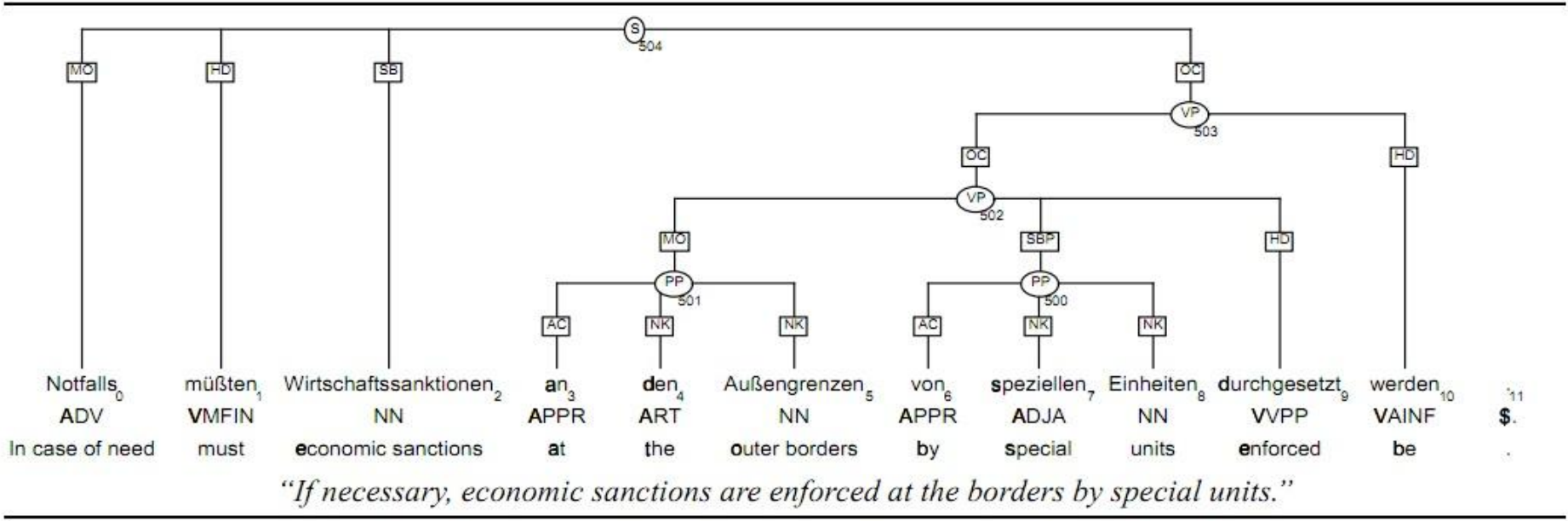

Figure 1: Example sentence with part-of-speech and structural annotation.

Das Problem & das Ziel

★ Das große Problem (Konsistenz zu erwarten):

- Was bedeutet das genau ?

- Zwei Annotators annotieren denselben Satz gleich! Das ist was man immer und bei jedem Annotator braucht. Weil die Konsistenz die Effektivität eines Korpus nicht nur fürs Testen, sondern auch für linguistische Untersuchungen braucht. Jedoch ist es aber aufwendig zu erhalten.

★ Warum Konsistenz?
- Um ein effektives Programm zu machen! Also, die Maschinen sollten alles allein machen.

Das Ziel:

Unterschiede zu entdecken bzw. klassifizieren. Damit wir die Konsistenz erhöhen können und die Fehler besser behandeln.

Eine Nebenwirkung ist höhere Annotationsgeschwindigkeit, seitdem weniger Unterschiede zu eliminieren sind.

Was haben wir benutzt bzw. getan?

Diese zwei Annotators haben wir genutzt, um eine gemeinsame finale Version daraus zu machen, und durch die Diskussion der Unterschiede und mehrfach wiederholungen von (Cleaning)Durchreiserung .

Die erste und die letzte Version sind verglichen worden. Daraus kamen zwei Kategorien:

- Die die größte Zahl von Unterschieden machten
- Die ab und zu auftauchend.

Die Strategie und die Maßnahmen

Während der Annotation vom NEGRA Korpus wurden effektive Werkzeuge entwickelt, die auf graphischem Feedback basiert.

Der Annotator wirkt mit Tagger und im Hitergrund laufendem Parser zusammen.

POS plus strukturelle Annotation .... Total Annotation ca. 10 min/Satz.Dies enthält:

1- Zwei voneinander unabhängige Annotationen.
2- Korrektur von offensichtlichen Fehlern während des Vergleichs.
3- Diskussion und Korrektur von den verbliebenen Unterschieden.

2&3 brauchen ziemlich viel Zeit, deswegen sind sie nicht zugestimmt.

4- Training phase.
5- Erforderliche Änderungen wegen dem Änderung bei dem Annotationsschema.

4& 5 brauchen feste Zeit, die aber von der ersten Phase des Annotators abhängt.

Maßnahmen

- Wir haben die initialen Versionen (A und B) und die finale(nach mehreren Schritten von Diskussion und Cleaning) Version miteinander verglichen.

- Dann benutzten wir die gleiche Maßnahme,die die Accuracy vom Tagging berechnet, für automatische Taggers(Ein Tag für jedes Wort).

Ø $$\text{Accuracy}(X,Y) = \frac{\text{number of tokens tagged identically}}{\text{number of tokens in the corpus}}$$

Wir könnten auch die standardmaßnahmen berechnen:

Recall: How much good stuff did we MISS!

Ø $$\text{Recall}(X,Y) = \frac{\text{number of identical nodes in X and Y}}{\text{number of nodes in X}}$$

Precision: How much junk are we giving to the user

Ø $$\text{Precision}(X,Y) = \frac{\text{number of identical nodes in X and Y}}{\text{number of nodes in Y}}$$

Ø F-Score (the harmonic mean of both Recall und Precision) $$F = \frac{2PR}{P+R}$$

Maßnahmen

Der F.score ist die meiste geeignete Maßnahme.

- ❏ Das Adaptieren den Denkansatz von Calder 1997:

Er benutzt terminal Produktion um Kontextfreie Bäume zuzuordnen.

- ❏ Das Erweitern:

zwei Knoten in zwei verschiedenen Annotationen sind identisch, wenn sie die gleichen terminal Produktion haben.
Very low F-Score identifiziert Kategorien, die durchgehend behandelt werden.

- Es war mehr oder weniger zufällig ob ein Annotator für Version A oder B oder zu einem Teil funktioniert hatte, deswegen stimmen unsere Ergebnisse miteinander manchmal ziemlich gut und manchmal ziemlich schlecht überein.

Results

- Die besten Ergebnisse für Kontextfreie Englisch Strukturen sind rund 86%. für die Deutsche Version 73%.
- Angenommen: Enthält unsere finale Version richtige Annotationen., heißt es, verringerung der Fehler auf 64% bis 81% von einer halb automatische Annotation.
- Für POS tags : Tags mit höherer Zahl von Unterschieden tendieren häufig zu sein und hoch F score zu haben.
- Auf der anderen Seite tendieren Tags mit sehr niedrigen F-Score unhäufig zu sein, jedoch ist cleaning Kategorien mit niedrigen F-scores sehr effektiv, wenn man sich für diese Untersuchungen interessiert.
- Maßnahmen, accuracy zu determinieren automatic tagging and parsing systems könnten auch für semi auomatic annotation gemacht werden.
- Die Zustimmung für die menschlichen Annotationen sind viel höher als die anderen
- Halb automatischer Pass reduziert die Fehlerquote auf 64_81% als bei den voll automatischen Vorgängen.
- Analysieren der Tags, die die große Unstimmigkeit machen, zeigt uns, dass höhere Unterschiede machende Kategorien stimmen nicht mit den Kategorien, die die höhere Zahl von Unterschieden machen.

Tabelle:(1)

Table 1: Agreement of part-of-speech annotations between two different annotators, and between the first and the final annotations.

Comparison	total number of tokens	agreement between A and B	FINAL and A	FINAL and B
Part-of-Speech	147,212	145,100 98.57%	145,445 98.80%	145,444 98.80%

Tabellen:(2) & (3)

Table 2: Part-of-speech tags which are involved in the highest numbers of differences when comparing annotations A and B. Note that the differences sum up to 200% since each difference involves two tags.

	tag	ident	diff	%total	F-score
1.	NN	31,331	670	31.7	98.9
2.	NE	7,553	580	27.5	96.3
3.	ADV	6,339	317	15.0	97.6
4.	ADJD	2,535	284	13.4	94.7
5.	ADJA	8,501	247	11.7	98.6
	— total —		4,224	200.0	98.6

Table 3: Part-of-speech tags with lowest F-scores when comparing annotations A and B. Note that the differences sum up to 200% since each difference involves two tags.

	tag	ident	diff	%total	F-score
1.	VMPP	1	3	0.1	40.0
2.	ITJ	6	10	0.5	54.6
3.	PTKANT	13	8	0.4	76.5
4.	FM	212	118	5.6	78.2
5.	PTKA	45	25	1.2	78.3
	— total —		4,224	200.0	98.6

Tabellen:(4)

Table 4: Pairs of part-of-speech tags with highest confusion rates when comparing annotations A and B.

	tag_1	tag_2	$f_1 + f_2$	diff	%total
1.	NE	NN	39,503	455	21.5
2.	ADJD	ADV	9,154	105	5.2
3.	ADJA	NN	40,297	74	3.5
4.	FM	NE	8,090	68	3.2
5.	PIAT	PIDAT	972	68	3.2
	— total —			2,112	100.0

Tabellen:(5)

Table 5: Agreement of structural annotations between two annotators, and between the first and the final annotations.

	recall		precision		F-Score
A vs. B					
unlabeled	67850 / 72319	(93.82%)	67850 / 72478	(93.61%)	(93.72%)
labeled	66921 / 72319	(92.54%)	66921 / 72478	(92.33%)	(92.43%)
incl. edge labels	64094 / 72319	(88.63%)	64094 / 72478	(88.43%)	(88.53%)
FINAL vs. A					
unlabeled	69646 / 73024	(95.37%)	69646 / 72319	(96.30%)	(95.84%)
labeled	68963 / 73024	(94.44%)	68963 / 72319	(95.36%)	(94.90%)
incl. edge labels	67273 / 73024	(92.12%)	67273 / 72319	(93.02%)	(92.57%)
FINAL vs. B					
unlabeled	69843 / 73024	(95.64%)	69843 / 72478	(96.36%)	(96.00%)
labeled	69183 / 73024	(94.74%)	69183 / 72478	(95.45%)	(95.10%)
incl. edge labels	67477 / 73024	(92.40%)	67477 / 72478	(93.10%)	(92.75%)

Tabellen:(6)&(7)

Table 6: Phrase types which are involved in the highest number of differences when comparing annotations A and B.

	phrase	ident	diff	F-score
1.	NP	19594	2996	92.9
2.	VP	5623	2294	83.1
3.	PP	17863	1705	95.4
4.	S	13477	1308	95.4
5.	AP	2371	898	84.1

Table 7: Phrase types with lowest F-scores when comparing annotations A and B.

	phrase	ident	diff	F-score
1.	CCP	1	2	50.0
2.	CO	41	64	56.2
3.	ISU	2	3	57.1
4.	DL	95	92	67.4
5.	AA	13	8	76.5

Der Schluss

Jetzt stellt sich die Frage, wie könnten wir das verbessern?

- Das nächste Analysieren ist in die Unterschiede mehr einzugehen, nach den Gründen von der höheren Zahl der Unterschiede zu suchen.
- Ob wir das Schema der Annotation ändern bzw. verbessern müssen oder nur zu den Annotators einfach mehr Training machen.