Thema:

Inter-Annotator Agreement for a German Newspaper Corpus

By:

Thorsten Brants

Where:

Saarland University, Computational Linguistics

Jahr: 2000

Inhaltverzeichnis:

Worüber geht es:

* Thema ?
* Das Problem
* Das Ziel.
* Die Strategie
* Die Ergebnisse.

Abstact:

Was hama ?

Zwischenkommentator

Agreement: Zustimmung oder Einigung

Untersuchung in

NEGRA Korpus, das aus Texten deutscher Zeitungen besteht. Das ist syntaktisch annotiert und strukturelle Information .

Zustimmung für POS ist 98.6%

By F- Score 92.4%

Recall

How much good stuff did we MISS!

Precise:

Proportion of retrieved set that are in fact relevant.

How much junk are we giving to the user

Diese zwei Annotators haben wir benutzt, um gemeinsame finale Version daraus zu machen, und zwar durch die Diskussion der Unterschiede und mehrfach wiederholungen von Durchreiserung . Die erste und die letzte Version sind verglichen worden. Daraus haben wir zwei Kategorien identifizert, die die großte Zahl von Unterschieden machten und auch Kategorien der ab und zu auftauchenden.

Consistency: Fogerichtigkeit, Konsistenz

Einführung:

Das große Problem (Konsistenz zu erwarten):

Was beduetet das genau ?  
zwei Annotator 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 wie auch für linguistische Untersuchungen. Jedoch ist es aber aufwendig zu erhalten.

Ein effektives verbessertes Werkzeug, das auf graphische feedback basiert,

Wirkt der Annotator mit Tagger und im Hitergrund laufenden Parser zusammen.

POS plus strukturelle Annotation .... Total Annotation ca. 10 min/Satz.

Das enthält:

1. Zwei voneinander unabhängige Annotationen.

Braucht weniger als zwei Minuten.

1. Korrektur von offensichtlichen Fehlern während des Vergleichs.
2. Diskussion und Korrektur von den verbliebenen Unterschiedlichen.

Brauchen ziemlich viel Zeit, deswegen sind sie nicht zugestimmt.

1. Training phase
2. Erfoderliche Änderungen wegen dem Änderung bei den Annotation Schema.

5& 6 brauchen feste Zeit, die aber von der ersten Phase der Annotator abhängt.

Das Ziel dieser Untersuchung ist die Unterschiede zu entdecken bzw. classifizieren.

Damit wir die Konsistenz erhöhen können und die Fehler besser behandeln zu können.

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

Wie

Untersuchungen sind gemacht an:

10,500 Sätze mit strukturellen aufgenommenen Änderungen und 8,500 Sätze mit aufgenommenen POS Annotationen Änderungen.

Beide wurden niedergeschreiben.

Andere Untersuchungen

* Voutilainen, 1999, der benutzte erfahrene Linguisten für die Annotation, während unsere Untersuchung von angestellten Studenten. Außerdem andere Tagset, die vermeidet manche Entscheidungen, die in unsere Tagset gemacht. Deswegen haben wir niedriger rates von Zustimmung erwartet
* Studie von Veronis 1998, beschäftigt sich mit Inter annotator agreement mit der Wortebene
* Mehrdeutlichkeit . POS die inofrmationen sind Wortebene annotiziert obwohl es komplett andere type von annotation ist.

2. Maßnahmen:/ Strategien

Results. Ergebnisse

POS

Structural Annotations:

Conclusions

Zusammenfassung:

Maßnahmen, accuracy zu determinieren automatic tagging and parsing systemen könnten auch für semi auomatic annotation gemacht werden.

Die Zustimmung für die menschlichen Annotationen sind viel höher als die anderen

Halb automatische pass reduziert die Fehlerquote auf 64\_81% als bei den voll automatischen Vorgängen.

Analysieren der Tags, die große Unstimmigkeit macht, zeigt uns, dass höhere Unterschiede machende Kategorien stimmen nicht mit den Kategorien, die höhreren Zahl von Unterschieden machen.

Die erste Neigung

Jetzt stellt sich die Frage … wie könnten

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.

Tabelle 1: Zeigt uns wie getaggt haben. zu jedem Wort wird ein Tag zugeordnet.

Aus den Versionen X und Y :

Accuracy (X,Y)- number of tokens tagged identically geteilt durch number of tokens in the corpus

Recall (X,Y) number of identical nodes in X and Y geteilt durch number of nodes in X

Precision (X,Y) number of identical nodes in X and Y geteilt durch number of nodes in Y

F-Score (the harmonic mean of both ) F 2PR geteilt durch P und R

F is eventuell die meisten geeignet Maßnahme, da noch keine identifizierte Version gibt.

Sehr niedrig F score identifiziert Kategorien mit inconsistency, die kommen zwar nicht häufig vor, allerdings erhöht die Behandlung von denen Corpus’s consistency