Text Data Mining – auf Grundlage von *Karine Megerdoomian (Handbook for Language Engineers)*

Seminar-Vortrag
im Rahmen der Computerlinguistik
Ausgearbeitet von
Daniel Gruber
(Juni 2004)

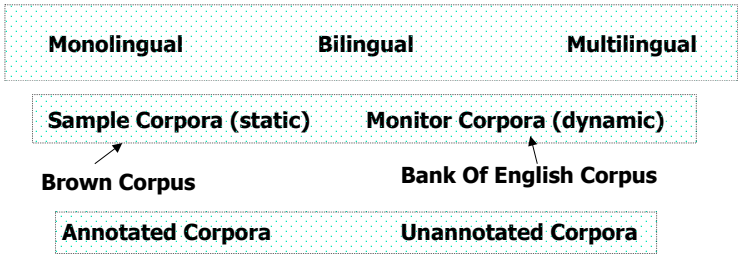
Corpus (Plural Corpora)

- A **collection** of written text or recorded speech
- Useful for *statistical* knowledge acquisition techniques
- Chomsky,1957,Syntactic Structures:
  „The corpus, if natural, will be so wildly skewed that the description [of language based on the corpus] would be no more then a mere list."

Corpus

- „A corpus is a *collection of text* or *speech* material that has been brought together according to be a certain set of predetermined criteria."
  - Testing hypotheses about natural lang.
  - Extracting statistical and linguistic information

Corpus – Types



Corpus consists of:
- → **text**
- → **corpus database that contains the text**
- → **concordancer as a front-end**

Corpus – Hitory

- Non-digital
  - Käding - 1897 – 100 million words [5000 analysts] → **shorthand/stenography**
  - Palmer – 1933 → language pedagogy
  - Eaton – 1940 → comparative linguistics
  - Fries, 1952 → corpus-based grammar
  - Quirk, 1961 → survey of English usage

Corpus – properties

- Machine readable form (digital)
- Representative of the domain under study
- Balanced sample
  - Text with specific parameters
  - (BNC, ANC)
- Finite (monitor corpus: non-finite [grows to reflect languages changes])

Corpora Resources I

- Brown Corpora (1 million tagged words)
  - Sample of written American English
- Lancaster-Oslo-Bergen corpus
- Penn Treebank
- British National Corpus (BNC)
- Project Gutenberg
  - http://www.promo.net/pg/
  - 6267 books free available
- Mannheimer Corpus Collection
  - http://corpora.ids-mannheim.de/cosmas/
  - (Demonstration einer statistischen Kookurenzanalyse)

Corpora Resources II

Brown Corpus	**Lancaster-Oslo-Bergen**	**Penn Treebank**	**British National Corpus (BE)**	**Birmingham Collection of English Text**	**Reuters**	**Project Gutenberg**	**American National Corpus**	**Mannheimer Corpus Collection**	**UN Parallel Text Corpus**
1967	1978		1994	1985	-	-			-
written AE	written BE		Written 90% spoken 10 %					German	English French Spanish
1.000.000 words		4.500.000 words	100.000.000 words	20.000.000 words	810.000 News	6267 Books		2 Bio. words	2.5 GB
		POS tagged	POS tagged						
		Wall Street Journal							UN electronic text archives (88-93)
									Multilingual

Corpus Analysis

- Corpus linguistics is the study of language through analysis of natural-occurring data. It involves computational methods and tools and develops theories of linguistics and language use.
- Annotation is mostly required for analyzing linguistic pattern
- Information Retrieval based on annotated corpora

Text Mining or Text Data Mining

- **Computer – Internet:** Enormous growth in the volume of online text documents in multiple formats and languages
- *Goal*: *discover knowledge* from unstructured textual data
- Text Mining vs. Data Mining
  - Type of data under investigation
  - TM: **unstructured** natural-language documents
  - DM: **highly structured** data in data warehouses
- Text Mining vs. Information Retrieval
  - DM: derive **new** information from data
  - IR: extracts already **existing** information

Text Mining – Process

(1) Information Extraction
  - TM systems include an IE module
  - Locates significant vocabulary items in NL documents (linguistic knowledge!)

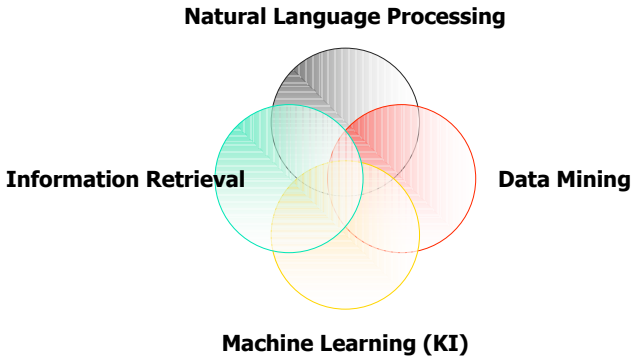
(2) Tokenization, stemming and tagging

(3) Cluster a collection of documents (IR)

(4) Categorize the clusters

(5) Discover knowledge from the databases and visualization tools (DM)

Text Mining – Text Data Mining



Tokenization – *pre-processing corpora I*

- *Problem one:* Text can be enriched with markups (HTML,XML,…), tables or other non-useful things…
- *Problem two:* We need a separated text to see **word and sentence boundaries**…
- Sentence Boundaries

? ! . Are good choices, but how can we handle an abbreviation (like Dr.) or acronym (like I.B.M.)? What's with Numbers (112.211)?

Haplology: One Character has two simultaneous uses. [i.e. the period at the end could signal both a sentence break and an abbreviation]

90% of the periods in English are sentence boundary markers.

Tokenization – *pre-processing corpora II*

- Sentence Boundaries

Solution: Regular Expressions with a list of abbreviations

Recent approaches:
Riley(1989): Statistical classification tree
Palmer and Hearst (1997): POS information used from a NN to predict sentence boundaries
Maximum entropy approach (probabilistic distribution of sentence boundaries in a text)

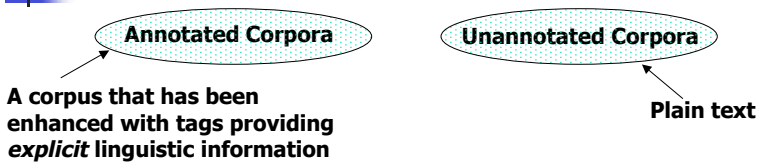
Tokenization – *pre-processing corpora III*

- Word Segmentation

-> Occurrence of whitespaces or punctuation
-> May work in English

- Punctuation: Apostrophes can be a part of a word (rock'n roll) or marking possessive nouns (cat's)
- Contractions: Spanish: del → de el
- Hyphenation: Compound word (e-mail) or joining words (25-year-old)
- Whitespace: New York
- More: Numbers, Encoding (→ Unicode)

Corpus Annotation



- How?
  - GML (Generalized Markup Language)
  - SGML (-> HTML)
  - XML (subset of SGML)
  - XCES (XML Corpus Encoding Standard [XML-scheme])
  - RDF (Resource Description Framework)

Corpus Annotation –
Annotation Coverage

Annotation schemes

- Part of speech annotation
  - One of the first types of annotation
  - Most common annotation today
- Lemmatization (or stemming)
  - Based on morphological analysis
- Parsing
  - Parsed corpora: *treebanks (→ Penn Treebank)*
  - Annotate the syntactic phrases (sentence, verb phrase, noun phrase, prepositional phrase)
- Semantic annotation, discourse analysis (*greetings, apologies [sorry], politeness [please]),* speech annotation

Corpus Annotation –
Tagset Design I

- Tagging is the foundation for further analysis
- Most common type is **part of speech (POS)** tagging
- Tagset: annotation tags used within the corpus
- Most widely used tagsets in:
  - Brown Corpus (http://www.scs.leeds.ac.uk/ccalas/tagsets/brown.html)
  - Penn Treebank (see next slide)
  - British National Corpus (BNC C5; distinguishes 61 categories; http://www.natcorp.ox.ac.uk/what/c5spec.html)

Corpus Annotation –
Tagset Design II

Penn Treebank Tagset

CC	Coordinating conjunction **e.g. and,but,or...**
CD	Cardinal Number
DT	Determiner
EX	Existential *there*
FW	Foreign Word
IN	Preposision or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List Item Marker
MD	Modal **e.g. can, could, might, may...**
NN	Noun, singular or mass
NNP	Proper Noun, singular

NNPS	Proper Noun, plural
NNS	Noun, plural
PDT	Predeterminer **e.g. all, both ... when they precede an article**
POS	Possessive Ending **e.g. Nouns ending in 's**
PRP	Personal Pronoun **e.g. I, me, you, he...**
PRP$	Possessive Pronoun **e.g. my, your, mine, yours...**
RB	Adverb Most words that end in -ly as well as degree words like quite, too and very
RBR	Adverb, comparative Adverbs with the comparative ending -er, with a strictly comparative meaning.
RBS	Adverb, superlative
RP	Particle
SYM	Symbol Should be used for mathematical, scientific or technical symbols
TO	*to*
UH	Interjection **e.g. uh, well, yes, my...**

http://www.computing.dcu.ie/~acahill/tagset.html

Corpus Annotation –
Tagset Design III

Penn Treebank Tagset

VB	Verb, base form subsumes imperatives, infinitives and subjunctives
VBD	Verb, past tense includes the conditional form of the verb to be
VBG	Verb, gerund or persent participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner **e.g. which, and** *that* when it is used as a relative pronoun
WP	Wh-pronoun **e.g. what, who, whom...**
WP$	Possessive wh-pronoun **e.g.**
WRB	Wh-adverb **e.g. how, where why**

Punctuation Tags
#
$
"
(
)
,
.
:
``

http://www.computing.dcu.ie/~acahill/tagset.html

Corpus Annotation –

Tagging Methods

- Rule-based taggers (knowledge-based taggers)
  - Lexicon based
  - TAGGIT (Green and Rubin, 1971)
  - Unable to provide tags for construction that have not been recognized
- Probabilistic taggers
  - Need to be trained to build a probability matrix (word/grammatical class/probability)
  - Bigram analysis (probability: a word of a certain POS follow a word from another particular POS) (trigrams)
  - (hidden Markov model -> viterbi algorithm)
  - If an unknown word occurs the grammatical class can be found thought the distributional information
- Hybrid taggers (CLAWS)

Corpus Annotation –

Tagging Methods: CLAWS

- BNC: 100.000.000 words
- Free WWW trail available http://www.comp.lancs.ac.uk/ucrel/claws/trial.html
- Tagsets: C5 (British National Corpus) and C7
- Output: horizontal or vertical (with columns)
- If you would like to use our trial service, please complete the form below. From an academic site, you can enter up to 10,000 words of English running text. From a non-academic site, you can enter up to 300 words of English running text. If you enter more, it will be cut off at the appropriate word limit. Input format guidelines are available. To tag the text you have entered click the button below the form.
- With C5 horizontal it looks like this:
- If_CJS you_PNP would_VM0 like_VVI to_TO0 use_VVI our_DPS trial_NN1 service_NN1 ,_, please_AV0 complete_VVB the_AT0 form_NN1 below_AV0 ._. From_PRP an_AT0 academic_AJ0 site_NN1 ,_, you_PNP can_VM0 enter_VVI up_AV021 to_AV022 10,000_CRD words_NN2 of_PRF English_AJ0 running_AJ0 text_NN1 ._. From_PRP a_AT0 non-academic_AJ0 site_NN1 ,_, you_PNP can_VM0 enter_VVI up_AV021 to_AV022 300_CRD words_NN2 of_PRF English_AJ0 running_AJ0 text_NN1 ._. If_CJS you_PNP enter_VVB more_AV0 ,_, it_PNP will_VM0 be_VBI cut_VVN off_AVP at_PRP the_AT0 appropriate_AJ0 word_NN1 limit_NN1 ._. Input_NN1 format_NN1 guidelines_NN2 are_VBB available_AJ0 ._. To_TO0 tag_VVI the_AT0 text_NN1 you_PNP have_VHB entered_VVN click_VVB the_AT0 button_NN1 below_PRP the_AT0 form_NN1 ._.

Applications in Corpus Linguistics

Extracting information

Training purposes

Applications in Corpus Linguistics –

Lexicon Acquisition

Extracting information

- Lexicon contains
  - Morphological
  - Syntactic
  - Semantic
  - Pragmatic information
- Lexicon used in
  - Information extraction
  - Document summarization
  - Machine translation
- Computational lexicon for taggers consists of the **lexemes** or **stem forms** of words
- If the tagger is used for a part of machine translation we need also the **translation**
- If incorrect information is in the lexicon the tagger cannot to be good as we want
- Machine readable dictionaries (MRD) containing lexical information
  - Static
  - Not available for many languages especially for translation purposes
- → Lexicon acquisition
  - Gather lexical information from text corpora
  - Corpora are now widely available
  - Can reflect dynamic and changing nature of language

Applications in Corpus Linguistics – Discourse Analysis

- Research topic: determining the psychological point of view (POV) of the author or a character in the text
- Current approaches:
  - Creating annotated corpora
  - Train a statistical discourse tagger module
  - Then the discourse tagger marks a corpus with POV expressions
  - → extremely difficult (no formal criteria)

Applications in Corpus Linguistics – Word Sense Disambiguation

- Goal: select the *appropriate meaning* to a given word based on the linguistic context
- Importance
  - → machine translation
  - → information retrieval
  - → parsing

Applications in Corpus Linguistics – Machine Translation

- Several methods (full knowledge-based systems, interlingua methods, purely statistical approaches,..)
- MT includes
  - Multilingual lexicon
  - Tagger
  - Parser
  - Word sense disambiguation module
- All are based on corpora
- MT specific: **parallel corpora** (same text in several languages)
- **Text alignment**: create explicit link between the elements that are mutual translations (→ *aligned corpus*)
- Methods for aligning sentences
  - Comparing the lengths of textual units
  - Using Lexical content
  - Matching cognates (verwandte)

Corpus Processing Tools

- UNIX/POSIX Tools
- Word Counts
- Concordances
- Collocations
- Testing and Evaluation

Corpus Processing Tools –
POSIX Tools

- POSIX: Portable Operating System Interface (...to be pronounced pahz-icks not poh-six)
- **POSIX Part 3: Shell and Utilities**
- **Standardization of tools so that they are widely available on several OS's.**
- A lot of utilities use *regular expressions* which are equivalent to *regular language* and equivalent to DFA (deterministic finite automata) and NFA (nondeterministic finite automata)
  - the word problem given from the expression can be solved efficiently
- Regular expressions are for instance:
  - A letter or a number
  - [...] := Class of characters – [^...] := complement class
  - Regular expression followed by +:= one or more
  - Regular expression followed by * := 0,1 or more (Kleene hull)
  - Regular expression | Regular expression := or
  - (Regular expression)
- Helmut Herold, Linux-Unix Kurzreferenz
- Regular expressions in C: man regex

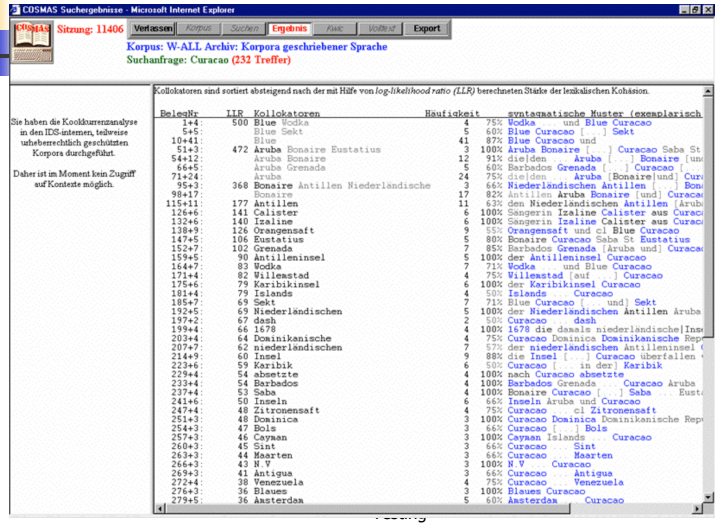
Corpus Processing Tools –
Word Counts

- Now we have some text on our POSIX-workstation
- We can use several utilities which are born to help us
- wc -w (for word count)
- tr (search and replace)
- cut (to handle text with columns)
- sort
- SED (stream editor) [sed 'script' file]
  - Script: [address1 [, address2]] function [args]
  - Addresses: number or /regex/
  - Functions: p (print), q (quit), s (replace)
  - sed '/Kommentar/!s/[^A-Za-z]/ /g' text.txt | wc -w
- AWK
  - Pattern { action}
  - Patterns: BEGIN, END, expression, regex, concatenated pattern, pattern1, pattern2 (all rows between pattern1 match and pattern2 match)
  - Builtin-variables: ARGC, ARGV, FILENAME, NF (number of fields)

Corpus Processing Tools –
Concordances

- Context of a word can be from interest
  - Especially for lexicographer
  - Intervening materials between verb and particle can be usefull for developing language grammars
- KWIC concordancing program (**K**ey **W**ord **i**n **C**ontext)
  - Extracts all occurences of the word of interest and displays it with the word in the center and the *surrounding context* on the two sides.

IDS Copora „Kookkurrenzanalyse" of „Curacao"



Corpus Processing Tools –

Collocations

- Concordance analysis is not very efficient for real quantitative analysis.
  - Not ordered by frequency
  - Number of hits can be very high
- „A *Collocation* is an **expression** that consists of a number of words within a short distance of each other."
- Collocation analysis
- Compositionality

Corpus Processing Tools –

Testing and Evaluation

- Test corpora
  - Corpora created to be used for evaluating and testing statistical algorithms and the performance of NLP systems.
  - Typically annotated
- Truthfile = annotated test corpus ?
- Split *data set* at the beginning into a *training set* and a *test set*
- Measurements:
  - Information Retrieval: Precision and Recall
  - Accuracy and coverage
- Criteria for a test set or truthfile
  - Consistency (reflect linguistic phenomena)
  - Size of Tagset (larger tagset → larger corpora to train sm)

Symbolic and Statistical Paradigms in Computational Linguistics

- Historically: conflict between symbolic and statistical approaches (Chomsky: Syntactic Structures (1957))
- Reasons
  - Role of quantitative measures
  - Type of data to be investigated
- Computational Approaches (in the 60s):
  - Generative theory; building linguistic models based on formal grammars
- 90s (computers faster; disk space cheaper):
  - Renewed interest in statistical models
- 1994: Workshop: The Balancing Act (New Mexico State University); Goal: dialogue between researchers of both sides
- Since then hybrid approaches growing

Summary

- Introduction to corpus linguistics
- Chomsky „Syntactic Structures" (1957)
- Corpus types and corpora resources
- Applications influences corpus collection and processing
- Tools and programs to analyse corpora

Discussion please…