

European Internet Accessibility Observatory

Introduction to Natural Language Processing

*HiA Webmining Course
2006-09-07*



Annika Nietzio
Forschungsinstitut Technologie-Behindertenhilfe (FTB)

About FTB ...

- Research institute
 - Technology for helping people with disabilities
- Accessibility
 - Public buildings
 - Independent living
 - Public transport
 - Assistive technology
 - Information technology / web accessibility

Outline

- What is language?
 - linguistic concepts
 - applications of computational linguistics
- Text mining and document classification
 - Vector space models
- Language modelling
 - Markov models

Outline

- *What is language?*
 - *linguistic concepts*
 - *applications of computational linguistics*
- Text mining and document classification
 - Vector space models
- Language modelling
 - Markov models

What is language?

When we study human language, we are approaching what some might call the “human essence”, the distinctive qualities of mind that are, so far as we know, unique to man.

(Noam Chomsky)

... what a famous linguist said.

What is language?

What you learn at school...

- Words
 - wordclasses
 - arbitrary relation of form and meaning
- Grammar
 - conjugation and declination
 - descriptive or prescriptive grammar

What is language?

Linguistic Layers

- Sounds
- Phonemes
- Words and Morphemes
- Syntax
- Semantics
- Text and Dialogue

Sounds and phonemes (1)

- Which sounds are part of a language?
 - more than 140 phonemes
 - phonetic system
 - vowels and consonants
- Minimal pair analysis
 - norwegian: lys – lus
 - englisch: coat – goat

Sounds and phonemes (2)

Let's do a little quiz.

`http://www.linguistics.ucla.edu/people/schuh/1x001/Web_Assignments/Assig_05/05web_04F.html`

Sounds and phonemes (3)

Application

- Speech recognition
 - Sound wave
 - Segmentation
 - Feature extraction
 - Classification (phonemes)
 - Mapping phonemes to letters / words
- Language model
- Which is the most likely phoneme sequence given the observation / recording?

Morphology (1)

- Morpheme = minimal unit of meaning
- Word formation
 - Compounds
 - web + mining
 - data + mining
 - web + mining + course
 - Derivation
 - friend + -ly + -ness
- Productivity
 - There are infinitely many (potential) words.

Morphology (2)

- Inflection
- Nouns:
 - cat – nominative singular
 - cat's – genitive singular
 - cats – nominative plural
- Verbs:
 - open – opens – opened – opening
- Stemming = remove inflectional endings

Words (1)

- Word classes, also called “Parts of Speech”
 - noun
 - verb
 - adjective
 - adverb
 - preposition
- Part of Speech Tagging
 - identify word class of each word in a sentence

Words (2)

- Mary saw the cat.
- Mary/**NE** saw/**V** the/**ART** cat/**NN** ./**\$**.
- Problem:
 - Ambiguity
 - The word 'saw' is used as verb not as noun.
 - Need to look at the whole sentence or at least a context.

Syntax (1)

- What is the structure of a sentence?
- Surface structure
 - We can only observe the word order.
- Grammar
 - rewrite rules [S \rightarrow NP VP]
 - symbols (terminal & non-terminal)

Syntax (2)

$S \rightarrow NP VP$

$NP \rightarrow NP PP \mid ART NN \mid NE$

$VP \rightarrow V \mid V NP \mid VP PP$

$PP \rightarrow P NP$

$ART \rightarrow \text{the}$

$NE \rightarrow \text{Mary}$

$NN \rightarrow \text{man} \mid \text{telescope}$

$P \rightarrow \text{with}$

$V \rightarrow \text{saw}$

Exercise: Use the grammar rules above to generate the sentence

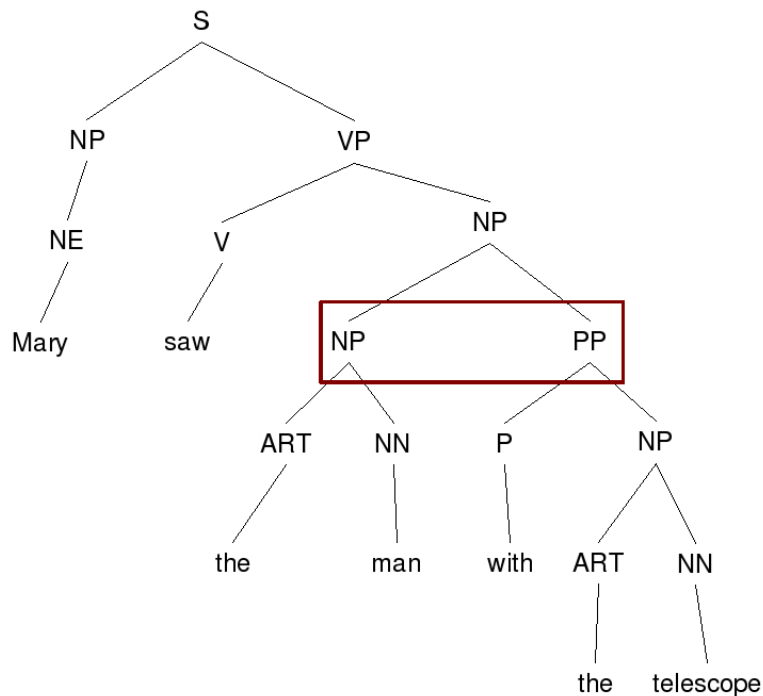
“Mary saw the man with the telescope”

Syntax (3)

There are two possible solutions.

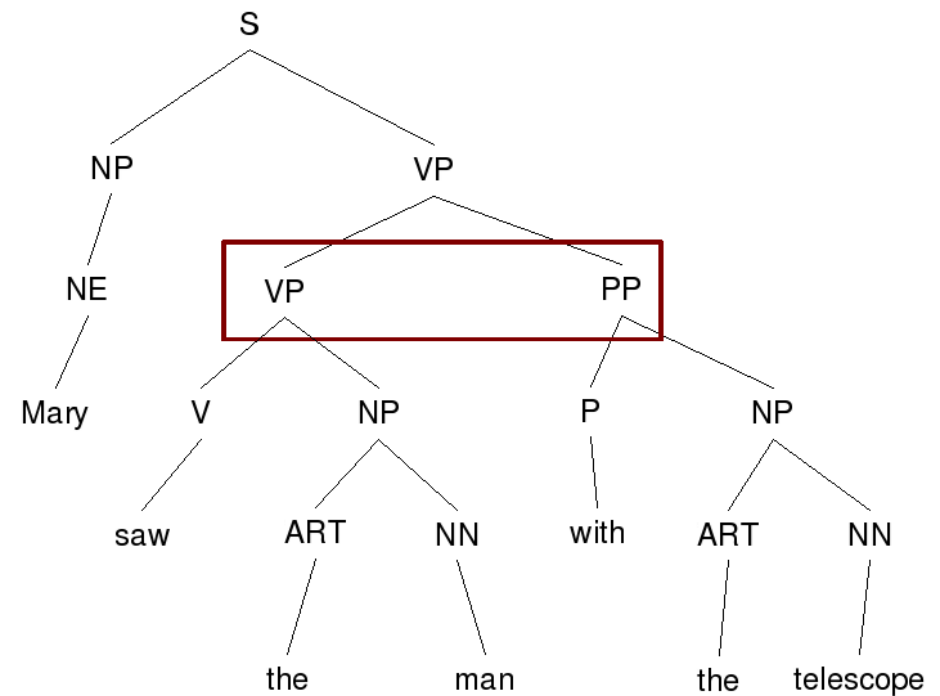
Solution A:

PP modifies the NP
(the man has the telescope)



Solution B:

PP modifies the VP
(Mary used the telescope)



Syntax (4)

Application

- Parsing
 - Natural language cannot be described by a regular grammar.
 - ... not even by a context-free grammar.
- Extension
 - Probabilistic context-free grammar
- Application
 - Determine if sentence is correct.
 - Prerequisite for semantic analysis, translation

Semantics (1)

- Meaning of language

What is the meaning of a word?

For a large class of cases – though not for all – in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in language.

(Wittgenstein)

Semantics (2)

- Logical representation
 - “Mary eats the cake.”
 - eat (Mary, cake)
 - “The cake eats Mary.”
 - Semantically not wellformed. Even if syntax is ok.
- Pragmatics
 - The intention of the speaker.
 - “It's cold in here.”
 - Meaning: “Please close the window.”

Semantics (3)

Application

- Word sense disambiguation
- Word senses (Example)
 1. note (noun) short piece of writing
 2. note (noun) a single sound at a particular level
 3. note (noun) a piece of paper money
 4. note (verb) to take notice of
 5. note (noun) of note: of importance
- Disambiguation via
 - Part of speech
 - Words in context

Conclusion (1)

- Language processing is difficult.
- Ambiguity in any layer.
 - Unlike programming languages, natural language is ambiguous if not understood in terms of all its parts.
 - Sometimes truly ambiguous.

Conclusion (2)

- Many theories and models.
- (Computing) resources are available.
- Tools involve
 - statistics
 - rule-based systems
 - logic
 - machine learning
- Language is produced with the intent of being understood.

Outline

- What is language?
 - linguistic concepts
 - applications of computational linguistics
- *Text mining and document classification*
 - *Vector space models*
- Language modelling
 - Markov models

Tasks and applications

- Document classification, text categorisation
 - emails (filtering, routing)
- Information retrieval, document retrieval
 - search engines
- Text mining
- Information extraction
 - from unstructured text (e.g. news) to database entries.
- Question answering systems
- Text summarisation

How to represent a document?

- We need a measure for the similarity of two documents.
- Indexing
 - Everything is indexed.
 - No predefined terms
 - Unlike databases
- Linguistic preprocessing
- Vector space model
 - also called “Bag of words”
 - Word order is not taken into account.

Linguistic preprocessing (1)

- Stop words
 - function words
 - occurring with high frequency in any text
- List of English stop words
 - a also an and as at be but by can could do for from go have he her here his how I if in into it its my of on or our say she that the their there therefore they this these those through to until we what when where which while who with would you your
- Remove stopwords
 - no relevant information about the document

Linguistic preprocessing (2)

- Stemming
 - Remove inflectional endings
 - Example: computing, computation, computable, computational are stemmed to *comput-*
- Useful for morphology-rich languages
- Problem
 - Semantically different words might be stemmed to the same root.

Vector space model (1)

- Introduced by Salton et al in 1975
 - *A vector space model for automatic indexing*
- Represent terms and documents as vectors.
 - dimensions = terms
 - values
 - 1 (present) or 0 (absent)
 - termfrequency $tf(i,j)$: number of occurrences of term i in document j
 - weighted termfrequency (to control the effect of the document length)

Vector space model (2)

Terms

1. Baby
2. Child
3. Guide
4. Health
5. Home
6. Infant
7. Proofing
8. Safety
9. Toddler

Documents

1. Infant & Toddler First Aid
2. Babies and Children's Room (for your Home)
3. Child Safety at Home
4. Your Baby's Health and Safety: From Infant to Toddler
5. Baby Proofing Basics
6. Your Guide to Easy Rust Proofing
7. Beanie Babies Collector's Guide

Vector space model (3)

Exercise: What is the term by document matrix for the example?

	D1	D2	D3	D4	D5	D6	D7
T1							
T2							
T3							
T4							
T5							
T6							
T7							
T8							
T9							

Vector space model (4)

The term by document matrix

	D1	D2	D3	D4	D5	D6	D7
T1	0	1	0	1	1	0	1
T2	0	1	1	0	0	0	0
T3	0	0	0	0	0	1	1
T4	0	0	0	1	0	0	0
T5	0	1	1	0	0	0	0
T6	1	0	0	1	0	0	0
T7	0	0	0	0	1	1	0
T8	0	0	1	1	0	0	0
T9	1	0	0	1	0	0	0

Similarity measures (1)

- Geometric properties in vector spaces
 - distances
 - angles
- Length of vector

$$\vec{x} = (x_1, \dots, x_n)$$

$$\|\vec{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$$

Euclidean distance:

$$d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|$$

Similarity measures (2)

Cosine similarity:

$$\cos(\vec{x}, \vec{y}) = \frac{\langle \vec{x}, \vec{y} \rangle}{\|\vec{x}\| \cdot \|\vec{y}\|}$$

- Meaning of the values
 - 1: angle 0 deg. Vectors point in same direction.
 - 0: angle 90 deg. Vectors are orthogonal.
 - 1: angle 180 deg. Vectors point in opposite directions.
- Usually used in information retrieval because it is insensitive to length of the term vectors.

Clustering the documents

Definition:

Grouping a collection of objects into subsets (*clusters*) such that those objects within each cluster are more closely related to one another than objects assigned to different clusters.

- Unsupervised learning
- Similarity measure to describe the relation of two objects

Document retrieval

- Given a query Q and a document D , what is the relevance of D to Q ?
- *Probability ranking principle*
 - underlying most IR systems
 - Rank documents in order of increasing probability of relevance is optimal.

Example (1)

- The normalised term by document matrix
 - Each column has length 1.
 - The normalised vector x' is computed as

$$\vec{x}' = \frac{1}{\|\vec{x}\|} \vec{x}$$

	D1	D2	D3	D4	D5	D6	D7
T1	0.000	0.577	0.000	0.447	0.707	0.000	0.707
T2	0.000	0.577	0.577	0.000	0.000	0.000	0.000
T3	0.000	0.000	0.000	0.000	0.000	0.707	0.707
T4	0.000	0.000	0.000	0.447	0.000	0.000	0.000
T5	0.000	0.577	0.577	0.000	0.000	0.000	0.000
T6	0.707	0.000	0.000	0.447	0.000	0.000	0.000
T7	0.000	0.000	0.000	0.000	0.707	0.707	0.000
T8	0.000	0.000	0.577	0.447	0.000	0.000	0.000
T9	0.707	0.000	0.000	0.447	0.000	0.000	0.000

Example (2)

Query “Child home safety”

Exercise:

- a) Calculate the (normalised) query vector.
- b) Calculate the cosine similarity for all documents.
- c) Which is the most relevant document for the query?

Terms:

1. Baby
2. Child
3. Guide
4. Health
5. Home
6. Infant
7. Proofing
8. Safety
9. Toddler

Solution

- Query vector
 - $q = (0, 1, 0, 0, 1, 0, 0, 1, 0)$
- Normalised query vector
 - $q' = (0, 0.577, 0, 0, 0.577, 0, 0, 0.577, 0)$
- Similarities
 - $qA = (0, 0.667, 1, 0.26, 0, 0, 0)$
- Most similar: D3

Search engines (1)

- Based on probability ranking principle.
- Additionally: more sophisticated ranking algorithms.
 - Page rank
 - Link counter
- Enhancement through user feedback
 - clickthrough data (record the selection of the user)

Search engines (2)

- Another strategy: manual indexing.
 - web directories
 - time and effort, much more expensive than automatic indexing
 - less consistent if several people are involved
- Main components of a search engine
 - crawler
 - indexing, analysis component
 - storage, databases
 - (fast) retrieval engine

Search engines (3)

If Only We Knew Google's Secret Ranking Algo ...

<http://blog.outer-court.com/archive/2006-07-31-n30.html>

Text Mining

“**Text Mining** is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.”

(Marti Hearst)

from <http://www.ischool.berkeley.edu/~hearst/textmining.html>

Performance evaluation (1)

- Precision
 - How many of the retrieved documents are really relevant?
 - Identifies relevant / irrelevant items.
- Recall
 - How many of the relevant document are retrieved from the collection?
 - Identifies included / excluded target items
- More sophisticated measures
 - take into account the ranking

Performance evaluation (2)

- Calculation of precision and recall requires human classification of documents to know which are “really relevant.”
- How can precision and recall be combined into one number?
 - Not independent. Not possible to just add them.
 - F-measure
 - typically $a = 0.5$

$$F = \frac{1}{a \frac{1}{P} + (1-a) \frac{1}{R}}$$

Outline

- What is language?
 - linguistic concepts
 - applications of computational linguistics
- Text mining and document classification
 - Vector space models
- *Language modelling*
 - *Markov models*

Modelling sequences

- Language is sequential.
 - utterance = sequence of phonemes
 - sentence = sequence of words
- Example: Part-of-Speech tagging
 - We want to find out
 - what is the Part-of-speech tag sequence for a given word sequence (sentence)?
 - PoS: short for Part-of-Speech

Statistical approach

- Ambiguity
 - Many words can have several PoS.
- Question (new version):
 - What is the *most likely* tag sequence for a given word sequence?

$$\operatorname{argmax}_{(t_1, \dots, t_n)} P(t_1, t_2, \dots, t_n | w_1, w_2, \dots, w_n)$$

where t_i is a tag, w_j is a word, and n is the length of the sentence

Hidden Markov Model (1)

- Given a sequence of random variables X with values from a finite set.
- X is called a **Markov Chain** if it has the following properties
 - **Limited horizon**

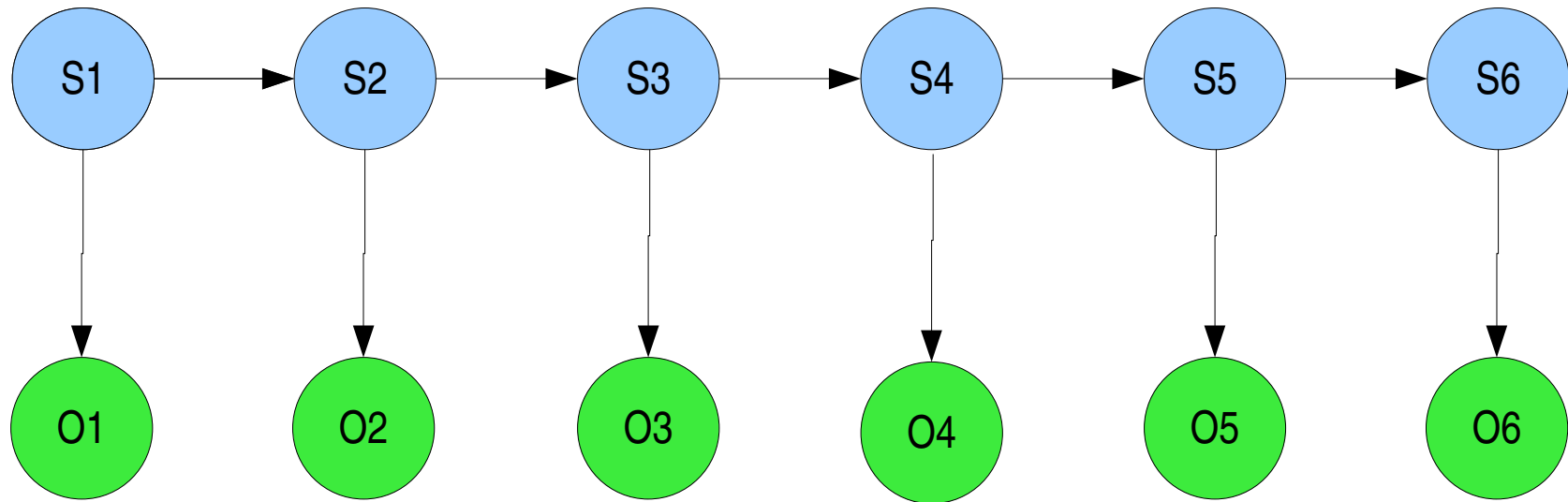
$$P(X_{t+1} = s_k | X_1, \dots, X_t) = P(X_{t+1} = s_k | X_t)$$

→ **Time invariance**

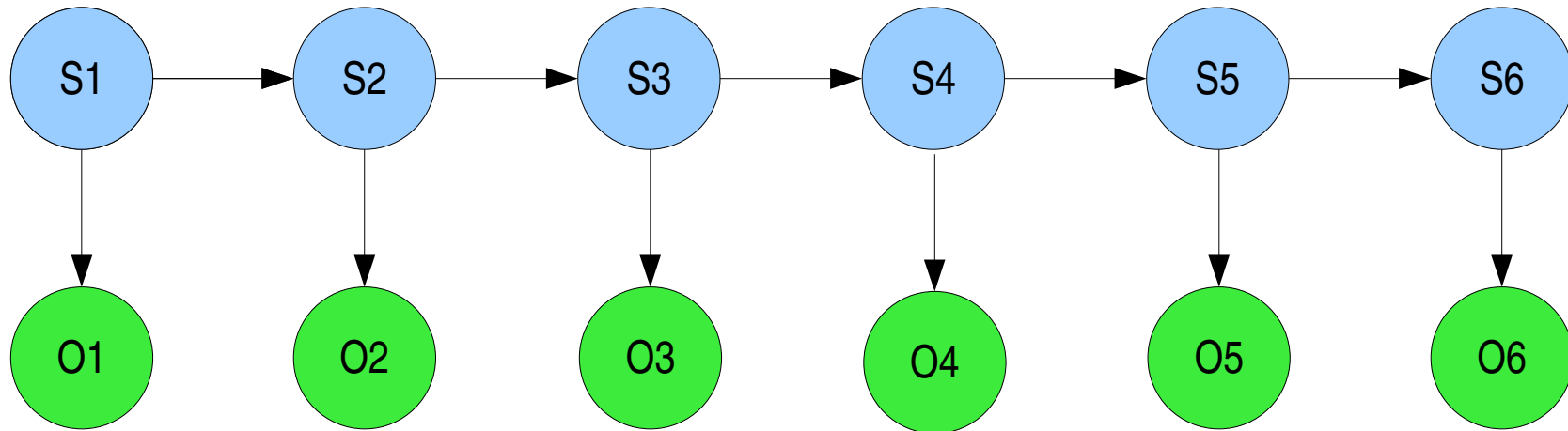
$$P(X_{t+1} = s_k | X_t) = P(X_2 = s_k | X_1)$$

Hidden Markov Model (2)

- The states of the model are hidden.
- We can only observe the symbols emitted at each stage.



Hidden Markov Model (3)



$$P(O_1, \dots, O_n | S_1, \dots, S_n) = P(S_1) P(O_1 | S_1) P(S_2 | S_1) P(O_2 | S_1, S_2) \dots \\ \dots P(S_n | S_{n-1}) P(O_n | S_{n-1}, S_n)$$

- Maximum likelihood estimate to determine probabilities. (Training)
- Dynamic programming algorithm for decoding. (Model application)

PoS Tagging (1)

- Determine the most likely tag sequence for a given word sequence.
- Mary/**NE** saw/**V** the/**ART** cat/**NN** ./**\$**.

$$\operatorname{argmax}_{(t_1, \dots, t_n)} P(t_1, t_2, \dots, t_n | w_1, w_2, \dots, w_n)$$

- Available: Annotated sentences for training.
 - Supervised machine learning.

PoS Tagging (2)

- Apply Bayes' rule:

$$\begin{aligned} & \operatorname{argmax}_{(t_1, \dots, t_n)} P(t_1, \dots, t_n | w_1, \dots, w_n) \\ &= \operatorname{argmax}_{(t_1, \dots, t_n)} \frac{P(w_1, \dots, w_n | t_1, \dots, t_n) P(t_1, \dots, t_n)}{P(w_1, \dots, w_n)} \\ &= \operatorname{argmax}_{(t_1, \dots, t_n)} P(w_1, \dots, w_n | t_1, \dots, t_n) P(t_1, \dots, t_n) \end{aligned}$$

- Hidden Markov model
 - words = observations
 - tags = states

Pos Tagging (3)

- n-grams
 - consider horizon $n > 2$
 - PoS-Tagging works best with 3-grams.
 - Sparse data problem for $n > 3$
- Smoothing
 - $P = 0$ for n-grams not appearing in the training data.
 - $P = 0$ for whole sentence.
 - Assign small probability to unobserved n-grams.

PoS Tagging (4)

- Performance:
 - 3-gram Taggers: ca. 95% correct tags
 - Baseline (most frequent tag): ca. 70%
 - Human experts agree on ca. 98% of tags.

Another application of the n-gram model

- Word prediction software.
- Tippfixx (developed at FTB)
 - For people with low input rate (keyboard or other device).
 - System tries to predict next word based on previous words and letters already typed.
 - User can select from list.
 - Saves 40-50% of keystrokes.

Summary

- Language processing
 - Linguistic layers building on top of each other.
 - Different sources of ambiguity
 - Natural language processing applied at each layer.
 - Automatic processing is possible.
- Tasks in information retrieval
 - Documents represented in vector space model
 - Document retrieval (example: search engine)
- Probabilistic language models
 - Sequences modelled by (Hidden) Markov models

References

- Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
- Daniel Jurafsky and James H. Martin, *SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2000.