Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Part-of-speech tagging

- Tagsets
- Constraint-based tagging
- Probabilistic tagger
  - Hidden Markov models
  - Maximum entropy models
- Transformation-based tagger
- Applications

Tagsets

- inventories of categories for the annotation of corpus data
- guided by lexical distributional classes
- but usually a more fine grained categorizations
  - morpho-syntactic subcategories (plural, tense, ...)
  - especially for the open classes: Nouns, Verbs, Adjectives and Adverbs
- inclusion of "technical" tags
  - foreign words, proper names, symbols, interpunction, ...

Tagsets

- typical tagsets

Penn-Treebank	Marcus et al. (1993)	45
British National Corpus (C5)	Garside et al. (1997)	61
British National Corpus (C7)	Leech et al. (1994)	146
Tiger (STTS)	Schiller, Teufel (1995)	54
Prague Treebank	Hajic (1998)	3000/1000

Tagsets

- Penn-Treebank (Marcus, Santorini, Marcinkiewicz 1993)

CC	Coordinating conjunction	*and,but,or, ...*
CD	Cardinal Number	*one, two, three, ...*
DT	Determiner	*a, the*
EX	Existential *there*	*there*
FW	Foreign Word	*a priori*
IN	Preposition or subord. conjunction	*of, in, by, ...*
JJ	Adjective	*big, green, ...*
JJR	Adjective, comparative	*bigger, worse*
JJS	Adjective, superlative	*lowest, best*
LS	List Item Marker	*1, 2, One, ...*
MD	Modal	*can, could, might, ...*
NN	Noun, singular or mass	*bed, money, ...*
NNP	Proper Noun, singular	*Mary, Seattle, GM, ...*
NNPS	Proper Noun, plural	*Koreas, Germanies, ...*
NNS	Noun, plural	*monsters, children, ...*

Tagsets

- Penn-Treebank (2)

PDT	Predeterminer	*all, both, ... (of the)*
POS	Possessive Ending	*'s*
PRP	Personal Pronoun	*I, me, you, he, ...*
PRP$	Possessive Pronoun	*my, your, mine, ...*
RB	Adverb	*quite, very, quickly, ...*
RBR	Adverb, comparative	*faster, ...*
RBS	Adverb, superlative	*fastest, ...*
RP	Particle	*up, off, ...*
SYM	Symbol	+, *%, & ...*
TO	*to*	*to*
UH	Interjection	*uh, well, yes, my, ...*
VB	Verb, base form	*write, ...*
VBD	Verb, past tense	*wrote, ...*
VBG	Verb, gerund	*writing*
VBN	Verb, past participle	*written, ...*

Tagsets

- Penn-Treebank (3)

VBP	Verb, non-3rd singular present	*write, ...*
VBZ	Verb, 3rd person singular present	*writes, ...*
WDT	Wh-determiner	e.g. *which, that*
WP	Wh-pronoun	e.g. *what, whom, ...*
WP$	Possessive wh-pronoun	*whose, ...*
WRB	Wh-adverb	e.g. *how, where, why*
$	Dollar sign	*$*
#	Pound sign	*#*
``	left quote	“
''	right quote	”
(	left parantheses	*(*
)	right parantheses	*)*
,	comma	,
.	sentence final punct.	*., !, ?*
:	mid-sentence punct.	*:, ;, –, ...*

Tagsets

- example for a tagged utterance
  - before disambiguation

    Book/NN/VB that/DT/WDT flight/NN ./.
  - after disambiguation

    Book/VB that/DT flight/NN ./.

Tagsets

- Stuttgart-Tübingen Tagset (STTS)
  (Schiller and Teufel 1995)

ADJA	attributives Adjektiv	das *große* Haus
ADJD	adverbiales oder prädikatives Adjektiv	er fährt/ist *schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in* der Stadt, *ohne* mich
APPRART	Präposition mit Artikel	*im* Haus, *zur* Sache
APPO	Postposition	ihm *zufolge*, der Sache *wegen*
APZR	Zirkumposition rechts	von jetzt *an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das, ein, eine, ...*
CARD	Kardinalzahl	*zwei* Männer, im Jahre *1994*
FM	Fremdsprachliches Material	Es wird mit "*A big fish*" übersetzt
ITJ	Interjektion	mhm, ach, tja
ORD	Ordinalzahl	[der] neunte [August]
KOUI	unterordn. Konjunktion mit "zu" + Infinitiv	*um/anstatt* zu leben
KOUS	unterordnende Konjunktion mit Satz	*weil, dass, damit, wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichskonjunktion	*als, wie*

Tagsets

- Stuttgart-Tübingen Tagset (STTS)(2)

NN	normales Nomen	*Tisch*, *Herr*, das *Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativpronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativpronomen	*jener* Mensch
PIS	substituierendes Indefinitpronomen	*keiner, viele, man, niemand*
PIAT	attrib. Indefinitpron. ohne Determiner	*kein/irgendein* Mensch,
PIDAT	attrib. Indefinitpron. mit Determiner	ein *wenig* Bier, *beide* Brüder
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessivpronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein* Buch, *deine* Mutter
PRELS	substituierendes Relativpronomen	der Hund, *der*
PRELAT	attribuierendes Relativpronomen	der Mann, *dessen* Hund
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche* Farbe, *wessen* Hut
PWAV	adverbiales Interrogativ oder Relativpronomen	*warum, wo, wann, worüber*
PAV	Pronominaladverb	*dafür, deswegen, trotzdem*

Tagsets

- Stuttgart-Tübingen Tagset (STTS)(3)

PTKZU	"zu" vor Infinitiv	*zu* gehen
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	er kommt *an*, er fährt *rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am* schönsten, *zu* schnell
SGML	SGML Markup	`<turnid=n022k_TS2004>`
SPELL	Buchstabierfolge	*S-C-H-W-E-I-K-L*
TRUNC	Kompositions-Erstglied	*An-* und Abreise
VVFIN	finites Verb, voll	du *gehst*, wir *kommen* [an]
VVIMP	Imperativ, voll	*komm* !
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit "zu", voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	du *bist*, wir *werden*
VAIMP	Imperativ, aux	*sei* ruhig !
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*gekonn*t, er hat gehen *können*
XY	Nichtwort, Sonderzeichen enthaltend	*3:7, H2O, D2XW3*

Tagsets

- Stuttgart-Tübingen Tagset (STTS)(4)

$,	Komma	*,*
$.	Satzbeendende Interpunktion	*. ? ! ; :*
$(	sonstige Satzzeichen; satzintern	*- [ ] ( )*

- examples (Tiger corpus)

  Werden/VAFIN sie/PPER diesmal/ADV lachen/VVINF //,/$. kreischen/VVINF ?/$.

  Mehr/PIAT Zeit/NN wenden/VVFIN die/ART US-Bürger/NN nur/ADV für/APPR Arbeiten/NN und/KON Schlafen/NN auf/PTKVZ ./$.

Constraint-based tagging

- ENGTWOL, Helsinki University (Voutilainen 1995)
- two-step approach
  - assignment of POS-hypotheses: morphological analyzer (two-level morphology)
  - selection of POS-hypotheses (constraint-based)
- lexicon with rich morpho-syntactic information

```
("<round>"
  ("round" <SVO><SV> V SUBJUNCTIVE VFIN (@+FMAINV))
  ("round" <SVO><SV> V IMP VFIN (@+FMAINV))
  ("round" <SVO><SV> V INF)
  ("round" <SVO><SV> V PRES -SG3 VFIN (@+FMAINV))
  ("round" PREP)
  ("round" N NOM SG)
  ("round" A ABS)
  ("round" ADV ADVL (@ADVL)))
```

Constraint-based tagging

- 35-45% of the tokens are ambiguous: 1.7-2.2 alternatives per word form
- hypothesis selection by means of constraints (1100)
  - linear sequence of morphological features
- example
  - input: *a reaction* **to** *the ringing of a bell*
  - dictionary entry:

```
("<to>"
  ("to" PREP)
  ("to" INFMARK> (@INFMARK>))
```

Constraint-based tagging

- example
  - constraint

```
("<to>" =0 (INFMARK>) (NOT 1 INF)
                      (NOT 1 ADV)
                      (NOT 1 QUOTE)
                      (NOT 1 EITHER)
                      (NOT 1 SENT-LIM))
```

    Remove the infinitival reading if immediately to the right of *to* no infinitive, adverb, citation, *either, neither, both* or sentence delimiter can be found.

Constraint-based tagging

- start with the set of candidate tags from the dictionary
- remove tags until a fixed point is reached
- or until only a single tag remains

- if contraints cannot disambiguate further, preference rules can be applied, e.g. frequency-based heuristics

Constraint-based tagging

- evaluation on an annotated testset ("gold standard")
- if the tagger assigns exactly one tag to every input wordform

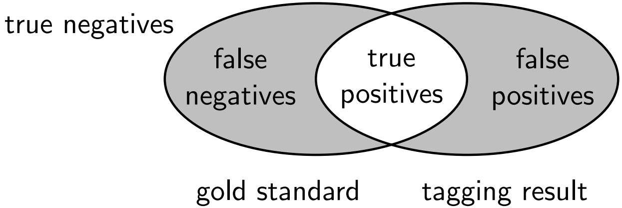
  $\rightarrow$ quality can be measured by means of accuracy

$$\text{accuracy} = \frac{\text{tags correctly assigned}}{\text{number of input wordforms}}$$

Constraint-based tagging

- if the tagging output is incomplete or ambiguous

  $\rightarrow$ quality needs to be measured by means of precision and recall



$$\text{recall} = \frac{|\text{true positives}|}{|\text{true positives}| + |\text{false negatives}|} = \frac{|\text{true positives}|}{|\text{gold standard}|}$$

$$\text{precision} = \frac{|\text{true positives}|}{|\text{true positives}| + |\text{false positives}|} = \frac{|\text{true positives}|}{|\text{tagging result}|}$$

Constraint-based tagging

General case: information retrieval (no disambiguation)

- true positives and false negatives are independent
- recall < 1: target items have not been found
- precision < 1: non-target items have been found

Special case incomplete disambiguation: $|\text{gold standard}| < |\text{tagging result}|$

- recall $>$ precision
- recall < 1: erroneous classifications, some constraints too strong

Special case incomplete tag assignment: $|\text{gold standard}| > |\text{tagging result}|$

- recall $<$ precision
- precision < 1: no classification results, rule set is overconstrained

Special case full disambiguation: $|\text{gold standard}| = |\text{tagging result}|$

- recall $=$ precision $\rightarrow$ accuracy

Constraint-based tagging

- recall and precision are antagonistic measures under the condition of limited competence:
  - increasing precision reduces recall
  - increasing recall reduces precision

- recall and precision can be combined into a single number:

  F-measure

$$F_\beta = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

for $\beta = 1$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \qquad \text{(harmonic mean)}$$

Constraint-based tagging

- ENGTWOL:
  - testset: 2167 word form token
  - recall: 99.77 %
  - precision: 95.94 %

  $\rightarrow$ incomplete disambiguation

Constraint-based tagging

- How good are the results?
  1. upper limit: How good is the annotation?
     - 96-97% agreement between annotators (Marcus et al. 1993)
     - almost 100% agreement in case of negotiation (Voutilainen 1995)
  2. lower limit: How good is the classifier?
     - baseline:
       e.g. most frequent tag (unigram probability)
     - example: $P(\mathrm{NN}|\mathit{race}) = 0.98$ $P(\mathrm{VB}|\mathit{race}) = 0.02$
     - 90-91% accuracy (Charniak et al. 1993)

Constraint-based tagging

- manual compilation of the constraint set and the dictionary
    - expensive
    - error prone
- alternative: using supervised machine learning approaches
    - (semi-automatic) annotation of training data is relatively easy
    - first success story in natural language processing

Probabilistic tagger

- Hidden Markov models
- Maximum entropy models

Hidden Markov models

- noisy-channel model
  - mapping from word forms to tags
  - mapping is not deterministic (ambiguity)
  - "noise" of the channel depends on the context
- Markov model: probabilistic model with memory
  - weighted finite state automaton
  - memory is modelled by means of states and state transitions
  - horizon is limited
  - transition probabilities $P(t_i|t_1 \ldots t_{i-1})$ can be aggregated to probabilities of tag sequences $P(t_{1 \ldots n})$
  - cannot accomodate ambiguity: one-to-one mapping between states and tags
    $\rightarrow$ Markov model needs to be extended

Hidden Markov models

- Hidden Markov model: probabilistic mapping between states (tags) and observations (wordforms)
  - sequence of state transitions influences the observation sequence probabilistically
  - captured by additional emission probabilities: $P(o_j | s_1 \ldots s_{j-1})$

Hidden Markov models

- model information for HMM taggers
  - observations, observation sequences:
    - word forms w_i
    - word form sequences $w_{1...n}$
  - model states, state sequences:
    - tags t_i
    - tag sequences $t_{1...n}$
  - transition probabilities: $P(t_i|t_1 \ldots t_{i-1})$
  - emission probabilities: $P(w_j|t_1 \ldots t_{j-1})$

Hidden Markov models

- some HMM transition probabilities can be deliberately set to zero
    - they define a specific model topology
- some HMM transition probabilities are zero because they have not been observed during training
    - artifacts of data sparseness

Hidden Markov models

- classification: computation of the most probable tag sequence

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}} P(t_{1\ldots n}|w_{1\ldots n})$$

- using Bayes' Rule

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}} \frac{P(t_{1\ldots n}) \cdot P(w_{1\ldots n}|t_{1\ldots n})}{P(w_{1\ldots n})}$$

- probability of the word form sequence is constant for a given observation and therefore has no influence on the decision result

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}} P(t_{1\ldots n}) \cdot P(w_{1\ldots n}|t_{1\ldots n})$$

Hidden Markov models

- using the chain rule for probabilities

$$P(t_{1\ldots n}) \cdot P(w_{1\ldots n} \mid t_{1\ldots n})$$

$$= \prod_{i=1}^{n} P(t_i \mid w_1 t_1 \ldots w_{i-1} t_{i-1})$$

$$\cdot P(w_i \mid w_1 t_1 \ldots w_{i-1} t_{i-1} t_i)$$

Hidden Markov models

- 1st simplification: the word form only depends on the current tag

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}}$$

$$\prod_{i=1}^{n} P(t_i \mid w_1 t_1 \ldots w_{i-1} t_{i-1}) \cdot P(w_i \mid t_i)$$

- 2nd simplification: the current tag depends only on its predecessors (not on the observations!)

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}} \prod_{i=1}^{n} P(t_i \mid t_1 \ldots t_{i-1}) \cdot P(w_i \mid t_i)$$

Hidden Markov models

- 3rd simplification: the current tag depends only on its two predecessors
  - limited memory (Markov assumption): trigram model

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}} \prod_{i=1}^{n} P(t_i \mid t_{i-1} t_{i-2}) \cdot P(w_i \mid t_i)$$

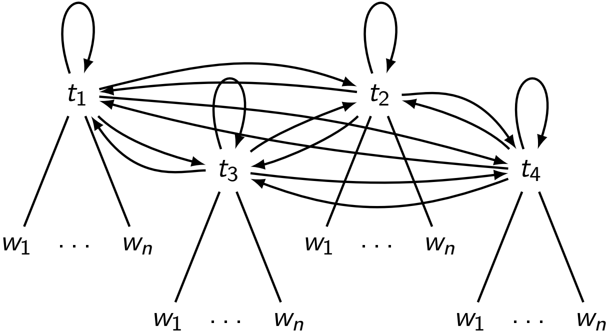
$\rightarrow$ 2nd order Markov process

Hidden Markov models

- further simplification leads to a bigram model
  - stochastic dependencies are limited to the immediate predecessor

$$\hat{t}_{1\ldots n} = \arg\max_{t_{1\ldots n}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \cdot P(w_i \mid t_i)$$

$\rightarrow$ 1st order Markov process



Hidden Markov models

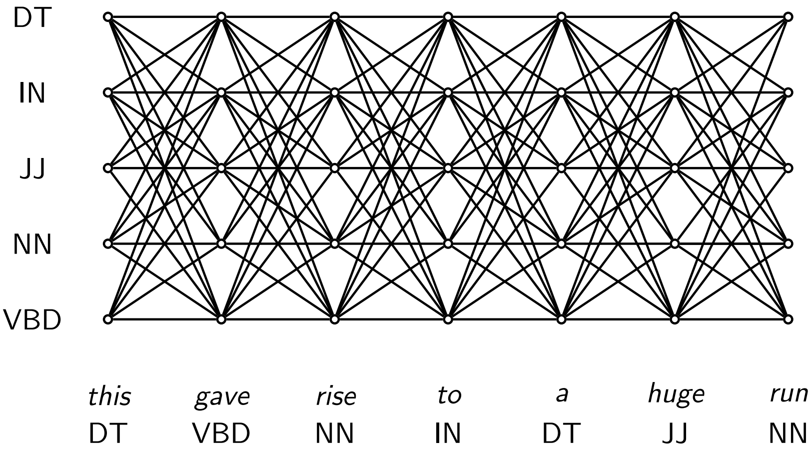
- decoding: computation of the most likely tag sequence
- using dynamic programming (Viterbi, Bellmann-Ford)

$$\delta(t_n) = \max_{t_{1\ldots n}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \cdot P(w_i \mid t_i)$$

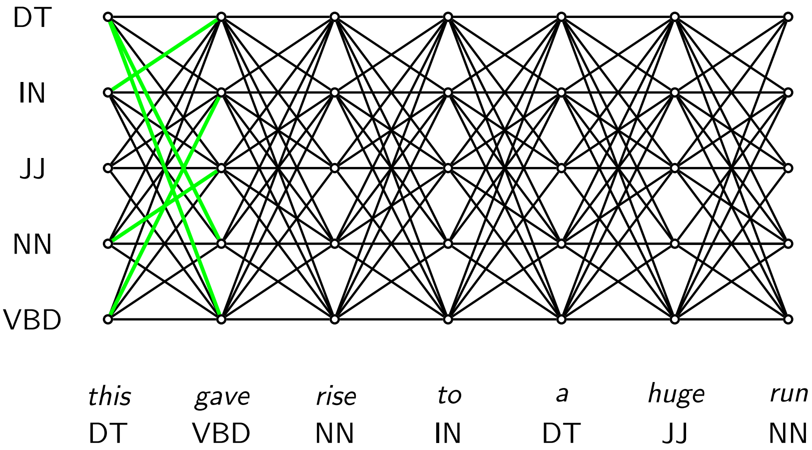
$$\delta(t_n) = \max_{t_{n-1}} \; P(t_n \mid t_{n-1}) \cdot P(w_n \mid t_n) \cdot \delta(t_{n-1})$$

- sometimes even local decisions are taken (greedy search)
- the tag sequence can be recovered by maintaining backpointers to the predecessor state which contributed the optimal path
- the scores can be interpreted as confidence values

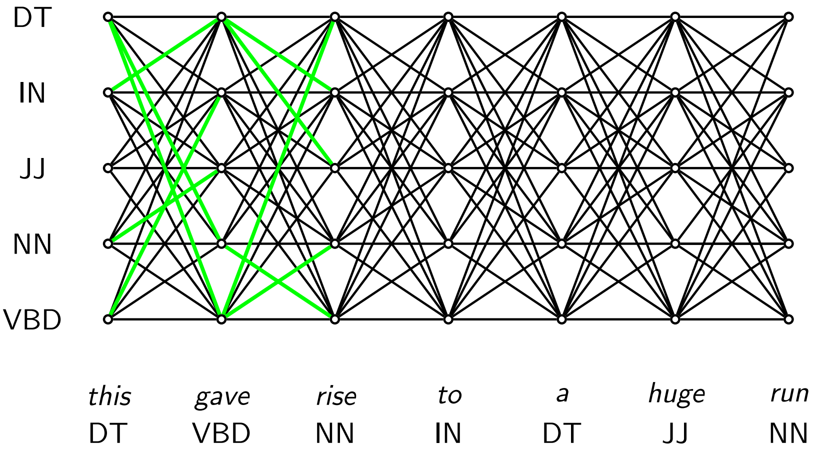
Hidden Markov models



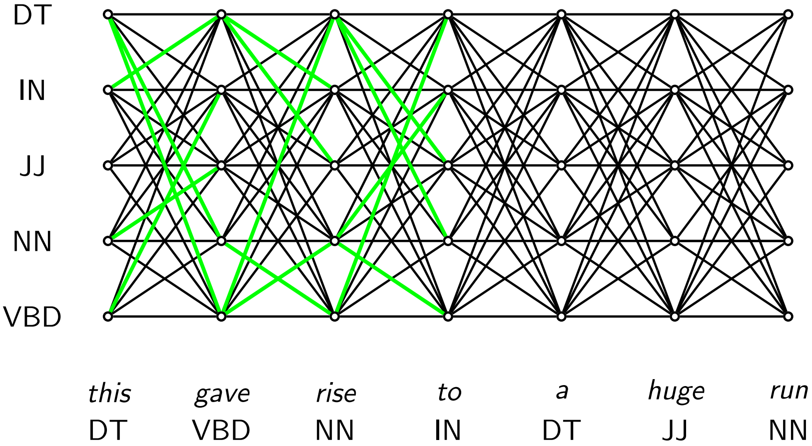
Hidden Markov models



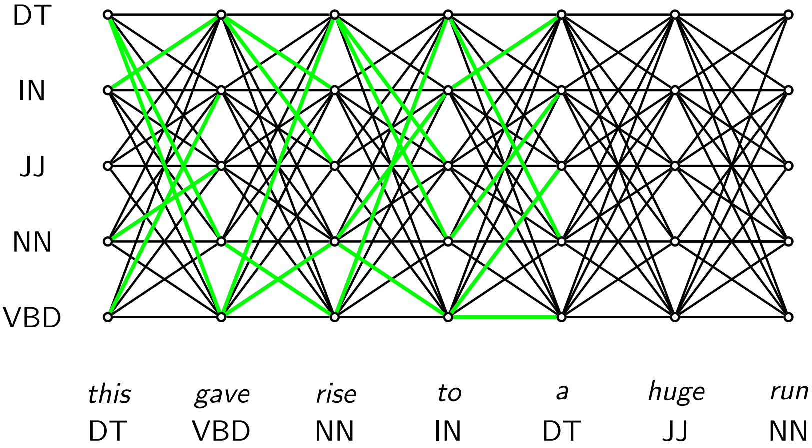
Hidden Markov models



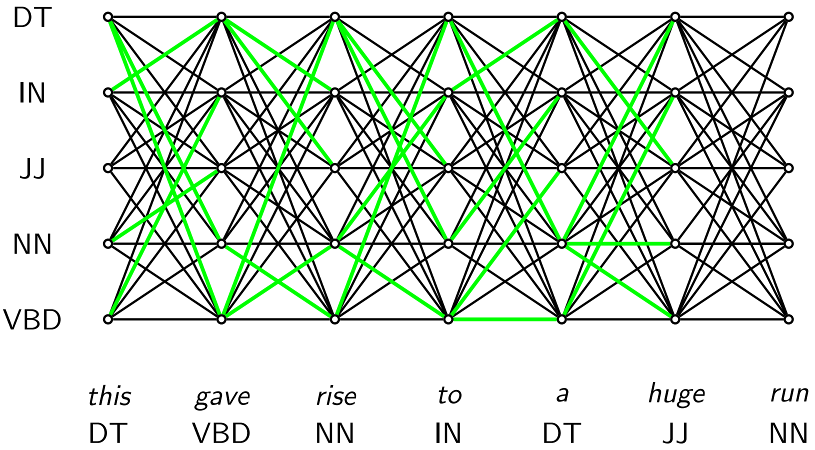
Hidden Markov models



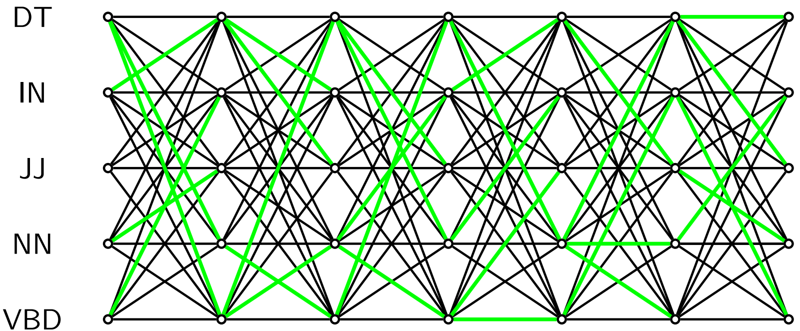
Hidden Markov models



Hidden Markov models

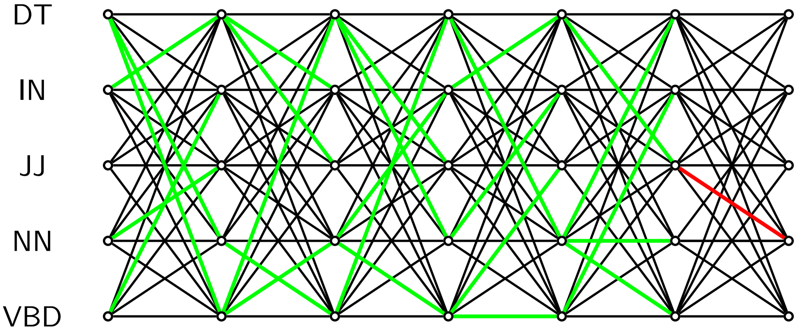


Hidden Markov models



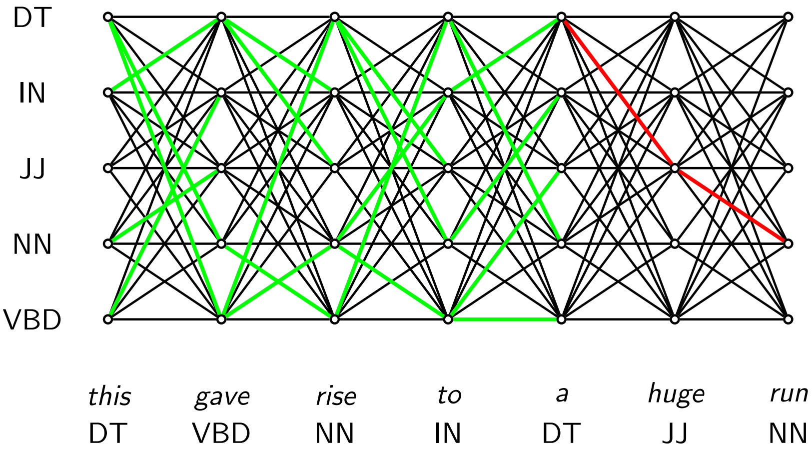
this	*gave*	*rise*	*to*	*a*	*huge*	*run*
DT	VBD	NN	IN	DT	JJ	NN

Hidden Markov models

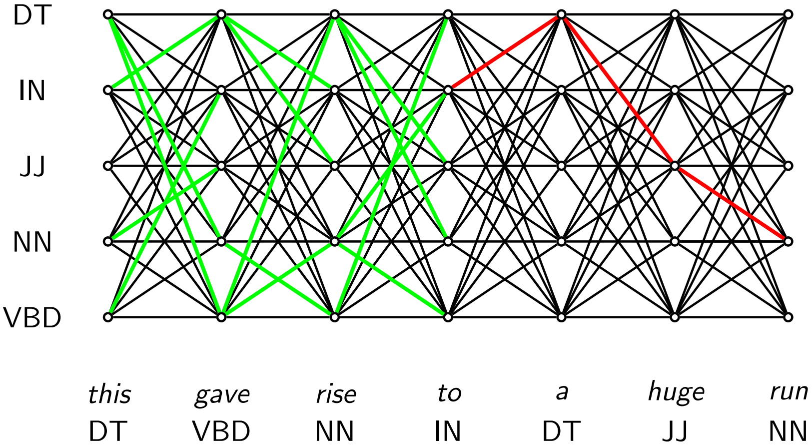


this	*gave*	*rise*	*to*	*a*	*huge*	*run*
DT	VBD	NN	IN	DT	JJ	NN

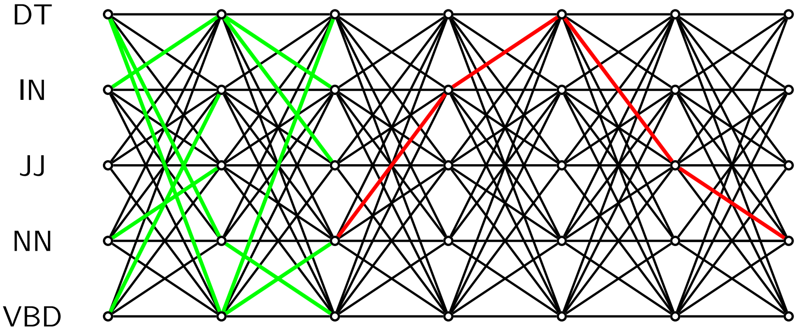
Hidden Markov models



Hidden Markov models

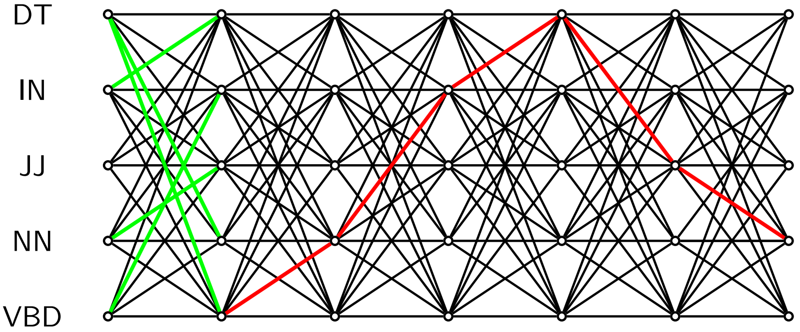


Hidden Markov models



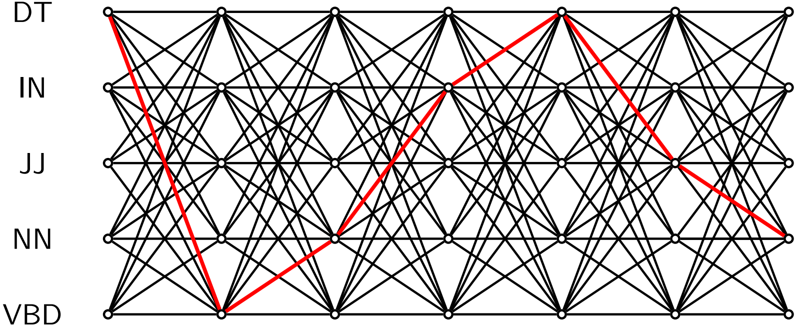
this	*gave*	*rise*	*to*	*a*	*huge*	*run*
DT	VBD	NN	IN	DT	JJ	NN

Hidden Markov models



this	*gave*	*rise*	*to*	*a*	*huge*	*run*
DT	VBD	NN	IN	DT	JJ	NN

Hidden Markov models



this *gave* *rise* *to* *a* *huge* *run*

DT VBD NN IN DT JJ NN

Hidden Markov models

- training: estimation of the probabilities on annotated corpus data
- fully observable data: maximum likelihood estimation can be applied
  - transition probabilities

$$P(t_i \mid t_{i-2}t_{i-1}) = \frac{c(t_{i-2}t_{i-1}t_i)}{c(t_{i-2}t_{i-1})}$$

  - emission probabilities

$$P(w_i \mid t_i) = \frac{c(w_i, t_i)}{c(t_i)}$$

Hidden Markov models

- unseen transition probabilities
  - backoff: using bigram or unigram probabilities in case the counts are too low (KATZ 1987)

$$P_{bo}(t_i|t_{i-2}t_{i-1}) = \begin{cases} \frac{c(t_{i-2}t_{i-1}t_i)}{c(t_{i-2}t_{i-1})} & \text{if} \quad c(t_{i-2}t_{i-1}t_i) > k \\ \lambda \cdot P_{bo}(t_i|t_{i-1}) & \text{else} \end{cases}$$

  - k is usually chosen to be zero
  - λ has to be determined on held out data (development set)

Hidden Markov models

- unseen transition probabilities
  - interpolation: linear combination of trigram, bigram, and unigram probabilities

$$P(t_i|t_{i-2}t_{i-1}) = \lambda_1 P(t_i|t_{i-2}t_{i-1}) + \lambda_2 P(t_i|t_{i-1}) + \lambda_3 P(t_i)$$

- λ_1, λ_2 and λ_3 are context dependent parameters
- global constraint: $\lambda_1 + \lambda_2 + \lambda_3 = 1$
- are trained on held out data (development set)

Hidden Markov models

- unseen word forms
  - estimation of the tag probability based on "suffixes" (and if possible also on "prefixes")
- unseen POS assignment
  - smoothing: redistribution of probability mass from the seen to the unseen events (discounting)
  - e.g. WITTEN-BELL discounting (WITTEN AND BELL 1991)
    - probability mass of the observation seen once is distributed to all the unseen events

Hidden Markov models

- tagging quality, TnT (Brants 2000)

corpus	unseen word-forms	accuracy known wordforms	accuracy unknown wordforms	overall
PennTB (English)	2.9%	97.0%	85.5%	96.7% (96.46%)
Negra (German)	11.9%	97.7%	89.5%	96.7%
Heise (German)*)				92.3%

*) test domain $\neq$ training domain

Hidden Markov models

- the standard output of a HMM tagger consists of
  - the optimal tag sequence
  - the probability/score of the optimal sequence

- multi-tagging: computing probability distributions for arbitrary tokens ("smoothing")

$$\alpha(t_n) = \sum_{t_{n-1}} P(t_n \mid t_{n-1}) \cdot P(w_n \mid t_n) \cdot \alpha(t_{n-1})$$

$$\beta(t_n) = \sum_{t_{n+1}} P(t_{n+1} \mid t_n) \cdot P(w_{n+1} \mid t_{n+1}) \cdot \beta(t_{n+1})$$

$$P(t_n) = \alpha(t_n) \cdot \beta(t_n)$$

Maximum entropy models

- Can we introduce additional wordform-related features into the decision?
  - capitalization (initial, middle, all), occurence of affixes, hyphens, digits, ...
- Can we directly train the conditional probability $P(t_{1...n}|w_{1...n})$?
  - HMM optimize the joint probability $P(t_{1...n}) \cdot P(w_{1...n}|t_{1...n})$
  - discriminative instead of generative models

$\rightarrow$ Maximum Entropy Models

Maximum entropy models

- multinomial logistic regression
    - linear regression
    - logistic regression
    - multinomial logistic regression
    - maximum entropy classifiers
    - maximum entropy Markov models

Linear regression

- representing the data as a combination of binary features
  - the (trigram) information for a sentence like

    This/DT gave/VBD rise/NN to/IN a/DT huge/JJ run/NN ./.

    can be encoded as

    $\langle$ VBD, w_{n-1} =this $\wedge$ t_{t-1}=DT $\rangle$
    $\langle$ NN, w_{n-2} =this $\wedge$ t_{t-2}=DT $\rangle$
    $\langle$ NN, w_{n-1} =gave $\wedge$ t_{t-2}=VDB $\rangle$
    ...
    $\langle$ NN, w_{n-2} =a $\wedge$ t_{t-2}=DT $\rangle$
    $\langle$ NN, w_{n-1} =huge $\wedge$ t_{t-2}=JJ $\rangle$
    ...

  - arbitrary other features can be added

Linear regression

- features are combined by a linear equation

$$y = w_0 + \sum_{i=1}^{n} w_i \cdot f_i$$

  - the weights describe how much influence a feature has on the decision to assign a particular tag

- the optimal set of weights can be found by minimizing the sum square error

$$e^2 = \sum_{t \in T} (y_{pred}(t) - y_{obs}(t))^2$$

  by means of a system of linear equations

Maximum entropy models

- the linear equations do not produce probabilities
- mapping to $[0,1]$ can be achieved by means of the logistic function

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

restricting y to {0,1} corresponding to {true,false}

$$P(y = true|x) = \frac{1}{1+e^{-\sum_i w_i \cdot f_i}}$$

$$P(y = false|x) = \frac{e^{-\sum_i w_i \cdot f_i}}{1+e^{-\sum_i w_i \cdot f_i}}$$

$\rightarrow$ logistic regression

Maximum entropy models

- training the logistic regression model by convex optimization techniques

$$\hat{W} = \arg\max_{W} \sum_{t \in T} y(t) \log \frac{1}{1 + e^{-\sum_i w_i \cdot f_i}}$$

$$+(1 - y(t)) \log \frac{e^{-\sum_i w_i \cdot f_i}}{1 + e^{-\sum_i w_i \cdot f_i}}$$

Maximum entropy models

- training the logistic regression model by convex optimization techniques

$$\hat{W} = \arg\max_{W} \sum_{t \in T} \overbrace{y(t) \log \frac{1}{1 + e^{-\sum_i w_i \cdot f_i}}}^{\text{positive samples}}$$

$$+ (1 - y(t)) \log \frac{e^{-\sum_i w_i \cdot f_i}}{1 + e^{-\sum_i w_i \cdot f_i}}$$

Maximum entropy models

- training the logistic regression model by convex optimization techniques

$$\hat{W} = \arg\max_{W} \sum_{t \in T} \overbrace{y(t) \log \frac{1}{1 + e^{-\sum_i w_i \cdot f_i}}}^{\text{positive samples}}$$

$$+ \underbrace{(1 - y(t)) \log \frac{e^{-\sum_i w_i \cdot f_i}}{1 + e^{-\sum_i w_i \cdot f_i}}}_{\text{negative samples}}$$

Maximum entropy models

- binary classification: assign true to a feature vector if

$$\sum_{i=0}^{n} w_i \cdot f_i > 0$$

- the linear equation

$$\sum_{i=0}^{n} w_i \cdot f_i = w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_n f_n = 0$$

  describes a hyperplane in the feature space, separating the positive from the negative cases

Maximum entropy models

- extension to multiple classes (multinomial logistic regression classification)
  - application of softmax to obtain probabilities

$$P(c|x) = \frac{e^{\sum w_{ci} \cdot f_i}}{\sum_{c' \in C} e^{\sum w_{c'i} \cdot f_i}}$$

  - choosing the class with the highest a posteriori probability

$$\hat{c} = \arg\max_c P(c|x)$$

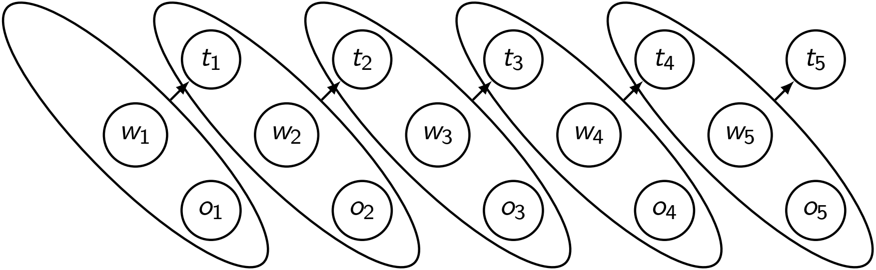
Maximum entropy models

- a model that assigns an even distribution to alternative feature values maximises entropy
    - but the training data constrain the possible combinations of feature values
- choose a model that assigns an equal probability to the alternative feature values given the constraints of the training data
    - adding features to the model selects subsets of training data that shall be modelled according to the empirical distribution in the training data
    - no other additional assumption shall be made
- such a model is equivalent to the probability distribution of a multinomial logistic regression model if the weights maximize the likelihood of the training data
- hence the name maximum entropy model

Maximum entropy models

- maximum entropy Markov models (MEMM)
- combining maximum entropy modelling with VITERBI search
    - only the probabilistic dependencies differ

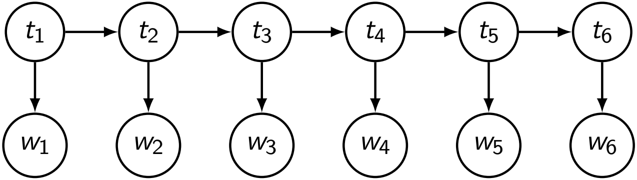
$$P(t_{1\dots n}) = \prod_{i=1}^{n} P(t_i | t_{i-1}, w_i, o_i, \dots)$$



Maximum entropy models

- HMM as a comparison

$$\begin{aligned} P(t_{1\ldots n}, w_{1\ldots n}) &= P(t_{1\ldots n}) \cdot P(w_{1\ldots n} | t_{1\ldots n}) \\ &= \prod_{i=1}^{n} P(t_i, w_i) \\ &= \prod_{i=1}^{n} P(t_i | t_{i-1}) \cdot P(w_i | t_i) \end{aligned}$$



Maximum entropy models

- MEMMs suffer from a label bias / observation bias
  - a strong source of evidence can explain away another one

example: *will/NN to/TO fight/VB*

- a modal verb after the lexical item *to* is less likely than a noun:

  $$P(TO|NN) > P(TO|MD)$$

  thus *will* should be tagged NN: *will/NN*

- the influence of the initial sentence position

  $$P(MD|\#will) > P(NN|\#will)$$

  together with a very strong influence of the lexical item *to*

  overrides this preference, because the previous tag does not really matter:

  $$P(TO|NN, to) \approx P(TO|MD, to) \approx P(TO|to)$$

Maximum entropy models

- the bias can be removed by
  - enriching the model with a bidirectional information flow
    e.g. Stanford tagger with cyclic dependency networks (MANNING 2011)
  - choosing a bi-directional model
    e.g. based on conditional random fields or
  - multiple pass tagging
- state-of-the art
  - bidirectional conditional random fields with long/short term memory and word embeddings as additional features (HUANG ET AL. 2015)

Maximum entropy models

- quality (Penn Treebank)

HMM	96.7% (96.46%)	Brants (2000)
ME classifier	96.6%	Ratnaparkhi (1996)
MEMM	96.96%	Denis and Sagot (2000)
ME cyclic dependencies	97.29%	Manning (2011)
CRF with l/s term memory	97.55%	Huang et al. (2015)

Transformation-based tagger

- ides: stepwise correction of wrong intermediate results (Brill 1995)
  - context-sensitive rules, e.g.
    Change NN to VB when the previous tag is TO
- rules are trained on a corpus
  1. initialisation: choose the tag sequence with the highest unigram probability
  2. compare the results with the gold standard
  3. generate a rule, which removes most errors
  4. run the tagger again and continue with 2.
- stop if no further improvement can be achieved

Transformation-based tagger

- rule generation driven by templates
  - change tag *a* to tag *b* if ...
    ...the preceding/following word is tagged *z*.
    ...the word two before/after is tagged *z*.
    ...one of the two preceding/following words is tagged *z*.
    ...one of the three preceding/following words is tagged *z*.
    ...the preceding word is tagged *z* and the following word is tagged *w*.
    ...the preceding/following word is tagged *z* and the word two before/after is tagged *w*.

Transformation-based tagger

- results of training: ordered list of transformation rules

from	to	condition	example
NN	VB	previous tag is TO	to/TO race/NN $\rightarrow$ VB
VBP	VB	one of the 3 previous tags is MD	might/MD vanish/VBP $\rightarrow$ VB
NN	VB	one of the 2 previous tags is MD	might/MD not reply/NN $\rightarrow$ VB
VB	NN	one of the 2 previous tags is DT	
VBD	VBN	one of the 3 previous tags is VBZ	

Transformation-based tagger

- 97.0% accuracy, if only the first 200 rules are used
- 96.8% accuracy with the first 100 rules
- quality of a HMM tagger on the same data (96.7%) is achieved with 82 rules
- extremly expensive training
  $\approx 10^6$ times of a HMM tagger

Applications

- word stress in speech synthesis
  'content/NN con'tent/JJ
  'object/NN ob'ject/VB
  'discount/NN dis'count/VB
- stemming, e.g. for document retrieval
- class based language models for speech recognition
- "shallow" analysis, e.g. for information extraction
- preprocessing for parsing data, especially in connection with data driven parsers

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Word sense disambiguation

- The problem
- Knowledge-based approaches
- Supervised Methods
- Semi-supervised Methods
- Baselines
- Unsupervised Methods

Word-sense disambiguation

"If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. ... But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word. ... The practical question is: 'What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?'"

Warren Weaver (1955)

"Sense ambiguity could not be resolved by electronic computer either current or imaginable."

Yoshua Bar-Hillel (1964)

Word-sense disambiguation

- word-sense disambiguation as a kind of tagging procedure?

Word-sense disambiguation

- word-sense disambiguation as a kind of tagging procedure?

- word senses are subjective
    - low degree of human agreement (80%)
    - even authoritative dictionaries do not agree
- word senses are difficult to deal with
    - they are difficult to imagine
    - they are difficult to memorize
- word senses are difficult to define
    - they are variable
    - the are not necessarily discrete
    - they are subject to gradual meaning shifts
    - there are problems of granularity and delineation

Word-sense disambiguation

- word-sense disambiguation as a kind of tagging procedure?
- word senses are context dependent
  - they can be modulated
  - they are task specific
- word senses behave fundamentally different from POS tags
  - way more "tags"
  - categories are to a large degree word specific
  - semantic influence is stretches across larger distances

Word-sense disambiguation

- How many senses has the noun *mark*? Which ones?

Word-sense disambiguation

- How many senses has the noun *mark*? Which ones?

1. S: (n) mark, grade, score (a number or letter indicating quality (especially of a student's performance)) "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?"

Word-sense disambiguation

- How many senses has the noun *mark*? Which ones?

1. S: (n) mark, grade, score (a number or letter indicating quality (especially of a student's performance)) "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?"
2. S: (n) marker, marking, mark (a distinguishing symbol) "the owner's mark was on all the sheep"

Word-sense disambiguation

- How many senses has the noun *mark*? Which ones?

1. S: (n) mark, grade, score (a number or letter indicating quality (especially of a student's performance)) "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?"
2. S: (n) marker, marking, mark (a distinguishing symbol) "the owner's mark was on all the sheep"
3. S: (n) target, mark (a reference point to shoot at) "his arrow hit the mark"

Word-sense disambiguation

- How many senses has the noun *mark*? Which ones?

1. S: (n) mark, grade, score (a number or letter indicating quality (especially of a student's performance)) "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?"
2. S: (n) marker, marking, mark (a distinguishing symbol) "the owner's mark was on all the sheep"
3. S: (n) target, mark (a reference point to shoot at) "his arrow hit the mark"
4. S: (n) mark, print (a visible indication made on a surface) "some previous reader had covered the pages with dozens of marks"; "paw prints were everywhere"

Word-sense disambiguation

- How many senses has the noun *mark*? Which ones?

1. S: (n) mark, grade, score (a number or letter indicating quality (especially of a student's performance)) "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?"
2. S: (n) marker, marking, mark (a distinguishing symbol) "the owner's mark was on all the sheep"
3. S: (n) target, mark (a reference point to shoot at) "his arrow hit the mark"
4. S: (n) mark, print (a visible indication made on a surface) "some previous reader had covered the pages with dozens of marks"; "paw prints were everywhere"
5. S: (n) mark (the impression created by doing something unusual or extraordinary that people notice and remember) "it was in London that he made his mark"; "he left an indelible mark on the American theater"

Word-sense disambiguation

6. S: (n) mark, stigma, brand, stain (a symbol of disgrace or infamy) "And the Lord set a mark upon Cain"–Genesis

Word-sense disambiguation

6. S: (n) mark, stigma, brand, stain (a symbol of disgrace or infamy) "And the Lord set a mark upon Cain"–Genesis
7. S: (n) mark, German mark, Deutsche Mark, Deutschmark (formerly the basic unit of money in Germany)

Word-sense disambiguation

6. S: (n) mark, stigma, brand, stain (a symbol of disgrace or infamy) "And the Lord set a mark upon Cain"–Genesis
7. S: (n) mark, German mark, Deutsche Mark, Deutschmark (formerly the basic unit of money in Germany)
8. S: (n) Mark, Saint Mark, St. Mark (Apostle and companion of Saint Peter; assumed to be the author of the second Gospel)

Word-sense disambiguation

6. S: (n) mark, stigma, brand, stain (a symbol of disgrace or infamy) "And the Lord set a mark upon Cain"–Genesis
7. S: (n) mark, German mark, Deutsche Mark, Deutschmark (formerly the basic unit of money in Germany)
8. S: (n) Mark, Saint Mark, St. Mark (Apostle and companion of Saint Peter; assumed to be the author of the second Gospel)
9. S: (n) chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug (a person who is gullible and easy to take advantage of)

Word-sense disambiguation

6. S: (n) mark, stigma, brand, stain (a symbol of disgrace or infamy) "And the Lord set a mark upon Cain"–Genesis
7. S: (n) mark, German mark, Deutsche Mark, Deutschmark (formerly the basic unit of money in Germany)
8. S: (n) Mark, Saint Mark, St. Mark (Apostle and companion of Saint Peter; assumed to be the author of the second Gospel)
9. S: (n) chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug (a person who is gullible and easy to take advantage of)
10. S: (n) mark (a written or printed symbol (as for punctuation)) "his answer was just a punctuation mark"

Word-sense disambiguation

11. S: (n) sign, mark (a perceptible indication of something not immediately apparent (as a visible clue that something has happened)) "he showed signs of strain"; "they welcomed the signs of spring"

Word-sense disambiguation

11. S: (n) sign, mark (a perceptible indication of something not immediately apparent (as a visible clue that something has happened)) "he showed signs of strain"; "they welcomed the signs of spring"
12. S: (n) Mark, Gospel According to Mark (the shortest of the four Gospels in the New Testament)

Word-sense disambiguation

11. S: (n) sign, mark (a perceptible indication of something not immediately apparent (as a visible clue that something has happened)) "he showed signs of strain"; "they welcomed the signs of spring"
12. S: (n) Mark, Gospel According to Mark (the shortest of the four Gospels in the New Testament)
13. S: (n) scratch, scrape, scar, mark (an indication of damage)

Word-sense disambiguation

11. S: (n) sign, mark (a perceptible indication of something not immediately apparent (as a visible clue that something has happened)) "he showed signs of strain"; "they welcomed the signs of spring"
12. S: (n) Mark, Gospel According to Mark (the shortest of the four Gospels in the New Testament)
13. S: (n) scratch, scrape, scar, mark (an indication of damage)
14. S: (n) crisscross, cross, mark (a marking that consists of lines that cross each other)

Word-sense disambiguation

11. S: (n) sign, mark (a perceptible indication of something not immediately apparent (as a visible clue that something has happened)) "he showed signs of strain"; "they welcomed the signs of spring"
12. S: (n) Mark, Gospel According to Mark (the shortest of the four Gospels in the New Testament)
13. S: (n) scratch, scrape, scar, mark (an indication of damage)
14. S: (n) crisscross, cross, mark (a marking that consists of lines that cross each other)
15. S: (n) bell ringer, bull's eye, mark, home run (something that exactly succeeds in achieving its goal) "the new advertising campaign was a bell ringer"; "scored a bull's eye"; "hit the mark"; "the president's speech was a home run"

Word-sense disambiguation

- How many senses has the verb *to mark*? Which ones?

Word-sense disambiguation

- How many senses has the verb *to mark*? Which ones?

1. S: (v) tag, label, mark (attach a tag or label to) "label these bottles"

Word-sense disambiguation

- How many senses has the verb *to mark*? Which ones?

1. S: (v) tag, label, mark (attach a tag or label to) "label these bottles"
2. S: (v) mark (designate as if by a mark) "This sign marks the border"

Word-sense disambiguation

- How many senses has the verb *to mark*? Which ones?

1. S: (v) tag, label, mark (attach a tag or label to) "label these bottles"
2. S: (v) mark (designate as if by a mark) "This sign marks the border"
3. S: (v) distinguish, mark, differentiate (be a distinctive feature, attribute, or trait; sometimes in a very positive sense) "His modesty distinguishes him from his peers"

Word-sense disambiguation

- How many senses has the verb *to mark*? Which ones?

1. S: (v) tag, label, mark (attach a tag or label to) "label these bottles"
2. S: (v) mark (designate as if by a mark) "This sign marks the border"
3. S: (v) distinguish, mark, differentiate (be a distinctive feature, attribute, or trait; sometimes in a very positive sense) "His modesty distinguishes him from his peers"
4. S: (v) commemorate, mark (celebrate by some ceremony or observation) "The citizens mark the anniversary of the revolution with a march and a parade"

Word-sense disambiguation

- How many senses has the verb *to mark*? Which ones?

1. S: (v) tag, label, mark (attach a tag or label to) "label these bottles"
2. S: (v) mark (designate as if by a mark) "This sign marks the border"
3. S: (v) distinguish, mark, differentiate (be a distinctive feature, attribute, or trait; sometimes in a very positive sense) "His modesty distinguishes him from his peers"
4. S: (v) commemorate, mark (celebrate by some ceremony or observation) "The citizens mark the anniversary of the revolution with a march and a parade"
5. S: (v) mark (make or leave a mark on) "the scouts marked the trail"; "ash marked the believers' foreheads"

Word-sense disambiguation

6. S: (v) stigmatize, stigmatise, brand, denounce, mark (to accuse or condemn or openly or formally or brand as disgraceful) "He denounced the government action"; "She was stigmatized by society because she had a child out of wedlock"

Word-sense disambiguation

6. S: (v) stigmatize, stigmatise, brand, denounce, mark (to accuse or condemn or openly or formally or brand as disgraceful) "He denounced the government action"; "She was stigmatized by society because she had a child out of wedlock"
7. S: (v) notice, mark, note (notice or perceive) "She noted that someone was following her"; "mark my words"

Word-sense disambiguation

6. S: (v) stigmatize, stigmatise, brand, denounce, mark (to accuse or condemn or openly or formally or brand as disgraceful) "He denounced the government action"; "She was stigmatized by society because she had a child out of wedlock"
7. S: (v) notice, mark, note (notice or perceive) "She noted that someone was following her"; "mark my words"
8. S: (v) scar, mark, pock, pit (mark with a scar) "The skin disease scarred his face permanently"

Word-sense disambiguation

6. S: (v) stigmatize, stigmatise, brand, denounce, mark (to accuse or condemn or openly or formally or brand as disgraceful) "He denounced the government action"; "She was stigmatized by society because she had a child out of wedlock"
7. S: (v) notice, mark, note (notice or perceive) "She noted that someone was following her"; "mark my words"
8. S: (v) scar, mark, pock, pit (mark with a scar) "The skin disease scarred his face permanently"
9. S: (v) score, nock, mark (make small marks into the surface of) "score the clay before firing it"

Word-sense disambiguation

6. S: (v) stigmatize, stigmatise, brand, denounce, mark (to accuse or condemn or openly or formally or brand as disgraceful) "He denounced the government action"; "She was stigmatized by society because she had a child out of wedlock"
7. S: (v) notice, mark, note (notice or perceive) "She noted that someone was following her"; "mark my words"
8. S: (v) scar, mark, pock, pit (mark with a scar) "The skin disease scarred his face permanently"
9. S: (v) score, nock, mark (make small marks into the surface of) "score the clay before firing it"
10. S: (v) set, mark (establish as the highest level or best performance) "set a record"

Word-sense disambiguation

11. S: (v) score, mark (make underscoring marks)

Word-sense disambiguation

11. S: (v) score, mark (make underscoring marks)
12. S: (v) cross off, cross out, strike out, strike off, mark (remove from a list) "Cross the name of the dead person off the list"

Word-sense disambiguation

11. S: (v) score, mark (make underscoring marks)
12. S: (v) cross off, cross out, strike out, strike off, mark (remove from a list) "Cross the name of the dead person off the list"
13. S: (v) check, check off, mark, mark off, tick off, tick (put a check mark on or near or next to) "Please check each name on the list"; "tick off the items"; "mark off the units"

Word-sense disambiguation

11. S: (v) score, mark (make underscoring marks)
12. S: (v) cross off, cross out, strike out, strike off, mark (remove from a list) "Cross the name of the dead person off the list"
13. S: (v) check, check off, mark, mark off, tick off, tick (put a check mark on or near or next to) "Please check each name on the list"; "tick off the items"; "mark off the units"
14. S: (v) grade, score, mark (assign a grade or rank to, according to one's evaluation) "grade tests"; "score the SAT essays"; "mark homework"

Word-sense disambiguation

11. S: (v) score, mark (make underscoring marks)
12. S: (v) cross off, cross out, strike out, strike off, mark (remove from a list) "Cross the name of the dead person off the list"
13. S: (v) check, check off, mark, mark off, tick off, tick (put a check mark on or near or next to) "Please check each name on the list"; "tick off the items"; "mark off the units"
14. S: (v) grade, score, mark (assign a grade or rank to, according to one's evaluation) "grade tests"; "score the SAT essays"; "mark homework"
15. S: (v) punctuate, mark (insert punctuation marks into)

Word-sense disambiguation

- lexical task
  - disambiguating selected words
  - e.g. *line-hard-serve* corpus: 4000 sense-tagged examples with *line* as a noun, *hard* as an adjective and *serve* as a verb
  - e.g. interest corpus: 2369 sense-tagged examples of *interest*
  - $\rightarrow$ training of supervised classifiers for individual words
- all-word task
  - using a corpus where *each* open class word is tagged with its actual sense
  - e.g. SemCor: 234.000 word subset of the Brown corpus tagged with WordNet senses
- human agreement: 75–80% for WordNet senses

Supervised methods

- use a (symmetric) window around the target word
- extract from the window
  - collocational features
    - position-oriented information about neighboring wordforms (wordform, base form, part-of-speech tag, . . .)
  - bag-of-word features
    - unordered set of neighboring wordforms
- train a supervised (word-specific) classifier
  - serious data sparsity problems

Pseudowords

- creation of artificial data for testing
    - concatenate two arbitrary words: e.g *bottle-cookie*
    - replace all occurrences of the two original words with the pseudoword
    - disambiguate the artificially introduced ambiguity using the original words as senses
    - compute disambiguation accuracy as usual
- evaluation with pseudowords is optimistic
    - natural sense alternatives are often similar
    - senses of pseudowords are not

Semi-supervised methods

- bootstrapping/boosting (Yarowski 1995)
  1. use a (small) sense-annotated seed corpus as training data
  2. train a classifier and classify unlabeled data
  3. add the high confidence results to the training set
  4. continue with 2

- heuristics for generating seed data
  - one sense per collocation: words with a strong association to the target sense tend to not occur with another sense
  - one sense per discourse: within a piece of text only one sense is used
  - selectional restrictions/preferences

Baselines

- take the most frequent sense
- Lesk algorithm(s):
  - compute the lexical overlap between the context of the target word and an entry in a dictionary or theaurus
  - simplified Lesk
    - choose the absolute number of common types
  - original Lesk
    - comparison with the context of context words
  - corpus Lesk
    - expanding the context of the target word with all the words that share the same sense in the annotated corpus
    - weighting the lexical overlap with the inverse document frequency

Unsupervised methods

- word-sense discrimination
- e.g. dynamic matching:
  - compare all instances of a given term in a corpus for common words and syntactic patterns
  - create a similarity matrix
  - cluster the words to find semantically related instances of the term
- results below the most-frequent sense baseline