

Empirical Approaches to Multilingual Lexical Acquisition

Lecturer: Timothy Baldwin



THE UNIVERSITY OF
MELBOURNE

Lecture 7

Learning Verb Syntax

Subcategorisation Frames

- “A subcategorisation (subcat) frame is a statement of what types of arguments a verb ... takes as objects, infinitives, *that*-clauses, participial clauses and subcategorised PPs” (Manning 1993):

John wants Mary to be happy

John hopes that Mary is happy

**John wants that Mary is happy*

**John hopes Mary to be happy*

Applications of Subcat Information

- Subcat information can lead to attachment disambiguation:

John put [the cactus] [on the table]

- Core component of type hierarchy in linguistically-precise grammars
- Empirical evidence for lexicalised subcat information improving the performance of statistical parsers, WSD systems, information extraction engines, etc.

From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax

(Brent 1993)

Basic Method

1. Identify verb tokens through a variety of heuristics
2. For each verb type, use high-precision lexico-syntactic patterns to identify evidence for 6 different subcat frames
3. Use a statistical filter to remove noise in the extracted subcat data

Identification of Verb Tokens

- Very rough and heuristic — (just) before the days of reliable POS tagging
- Focus on base and present participial verb forms
- Problems in distinguishing between base-form verbs and singular nouns (e.g. *record* — only workaround a filter on the immediately preceding word)

Lexico-syntactic Patterns

- Based on closed-class words (pronouns, determiners, complementisers, auxiliaries, punctuation)
- NPs captured in the form of pronouns or sequences of capitalised words
- VPs based on auxiliaries and the verbs learned in step 1

Statistical Filtering (1)

- Assumption that the probability of false evidence for a given subcat frame S (e.g. transitive) occurring is equal for all verbs incompatible with S (e.g. *snore*, *put*, *say*, ...)
- NOTE: probability of false evidence (π_{-S}) constant for a given S but varies across different subcat frames
- Null hypothesis: the verb does not belong to subcat class S , i.e. it is $-S$

Statistical Filtering (2)

- **Binomial test:** the probability of an event with probability p occurring exactly m out of n times is given by

$$P(m, n, p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

- The probability of the event occurring m or more times out of n is given by

$$P(m+, n, p) = \sum_{i=m}^n P(i, n, p)$$

$\frac{m}{n}$	$P(m, n, p = 0.1)$	$P(m+, n, p = 0.1)$
$\frac{0}{10}$	0.349	1.000
$\frac{1}{10}$	0.387	0.651
$\frac{2}{10}$	0.194	0.264
$\frac{3}{10}$	0.057	0.070
$\frac{4}{10}$	0.011	0.013
$\frac{5}{10}$	0.001	0.002
$\frac{6}{10}$	0.000	0.000
$\frac{7}{10}$	0.000	0.000
$\frac{8}{10}$	0.000	0.000
$\frac{9}{10}$	0.000	0.000
$\frac{10}{10}$	0.000	0.000

Statistical Filtering (3)

- Given n and p ($= \pi_{-S}$), we can apply a threshold θ to determine m such that verbs which occur with subcat frame S at least m times can be classified as $+S$ with $(1 - \theta)$ confidence
- In practice we don't know π_{-S} for each subcat frame S
SOLUTION: set θ and n , and estimate p based on the histogram distribution around each m ; select the p which best fits the binomial distribution

Shortcomings of the Brent Approach

- Assumption of π_{-S} being equal for all verbs given a class S shown to be flawed due to verb detection method
- Applicability of method to low-frequency words
- Scalability of method to other subcat frames

An Update on more Recent Research

- Greater coverage of subcat frames (up to 160)
- Simple frequency shown to be at least as effective as binomial test at filtering out noise
- Verb sense shown to interface closely with subcategorisation properties
- AND YET the Brent method still has remarkable currency to this day

Open Questions

- How to deal with low-frequency occurrences of subcat frames
- How well do the proposed methods port to other word classes (adjectives, nouns, ...) and languages
- Challenges for subcat acquisition in pro-drop languages (e.g. Japanese)

Alternations

(Baldwin and Bond 2002)

Definition of Alternation

- A regular mapping between argument positions in subcategorisation frames (generally assuming preservation of case-roles)
- Alternations involve *at least* one of:
 - i. word order/(prepositional, case, etc.) marking variation between corresponding case slots
 - ii. case slot deletion
 - iii. case slot insertion

Example English Alternations

- | | | |
|-----|--|-------------------|
| (1) | Kim loaded the truck with hay
Kim loaded hay on the truck | Spray/load |
| (2) | Kim sold the car to Sandy
Kim sold Sandy the car | Dative |
| (3) | The dog walks
Kim walks the dog | Causative |
| (4) | Kim sliced the meat
The meat sliced easily | Middle |

Example Japanese Alternations

- (5) Kim-ga doa-o akeru / doa-ga aku
Kim-NOM door-ACC opens door-NOM opens
'Kim opens the door' 'The door opens'
- (6) Kim-ga doa-o hiraku / doa-ga hiraku
Kim-NOM door-ACC opens door-NOM opens
'Kim opens the door' 'The door opens'
- (7) Kim-ga doa-o akeru / doa-ga ake-rareru
Kim-NOM door-ACC opens door-NOM opens-PASS
'Kim opens the door' 'The door is opened'

Types of Alternations (1)

- **Analytical/diathesis:** alternation unmarked on the verb (e.g. *hiraku* “open_{trans}” / *hiraku* “open_{intrans}”)
- **Lexical:** alternation marked on the verb stem by predictable lexical variation (e.g. *akeru* “open_{trans}” / *aku* “open_{intrans}”)
- **Synthetic:** alternation marked by verbal inflection or a verb morpheme (e.g. *taberu* “eat” / *tabe-saseru* “make eat”)

Types of Alternations (2)

- **Cognitive:** distinct verb forms but regularised pattern of alternation/simple change in focus, empathy, etc. (e.g. *kau* “buy” / *uru* “sell”)
- Focus on diathesis, lexical and synthetic in this research

Alternations and Verb Semantics

- Verbs with similar alternation behaviour shown to cluster together semantically
- Semantically-similar verbs shown to alternate similarly

- Example: verbs of contact:

- ★ CONATIVE alternation:

Kim punched the wall/Kim punched at the wall

- ★ BODY-PART POSSESSOR alternation:

Kim hit Sandy's finger/Kim hit Sandy on the finger

- ★ MIDDLE alternation:

Kim cut the bread/The bread cut easily

- ★ Verb classes:

Alternation	TOUCH	HIT	CUT	BREAK
CONATIVE	N	Y	Y	N
BODY-PART POSS	Y	Y	Y	N
MIDDLE	N	N	Y	Y

Alternation-based Lexicon Reconstruction

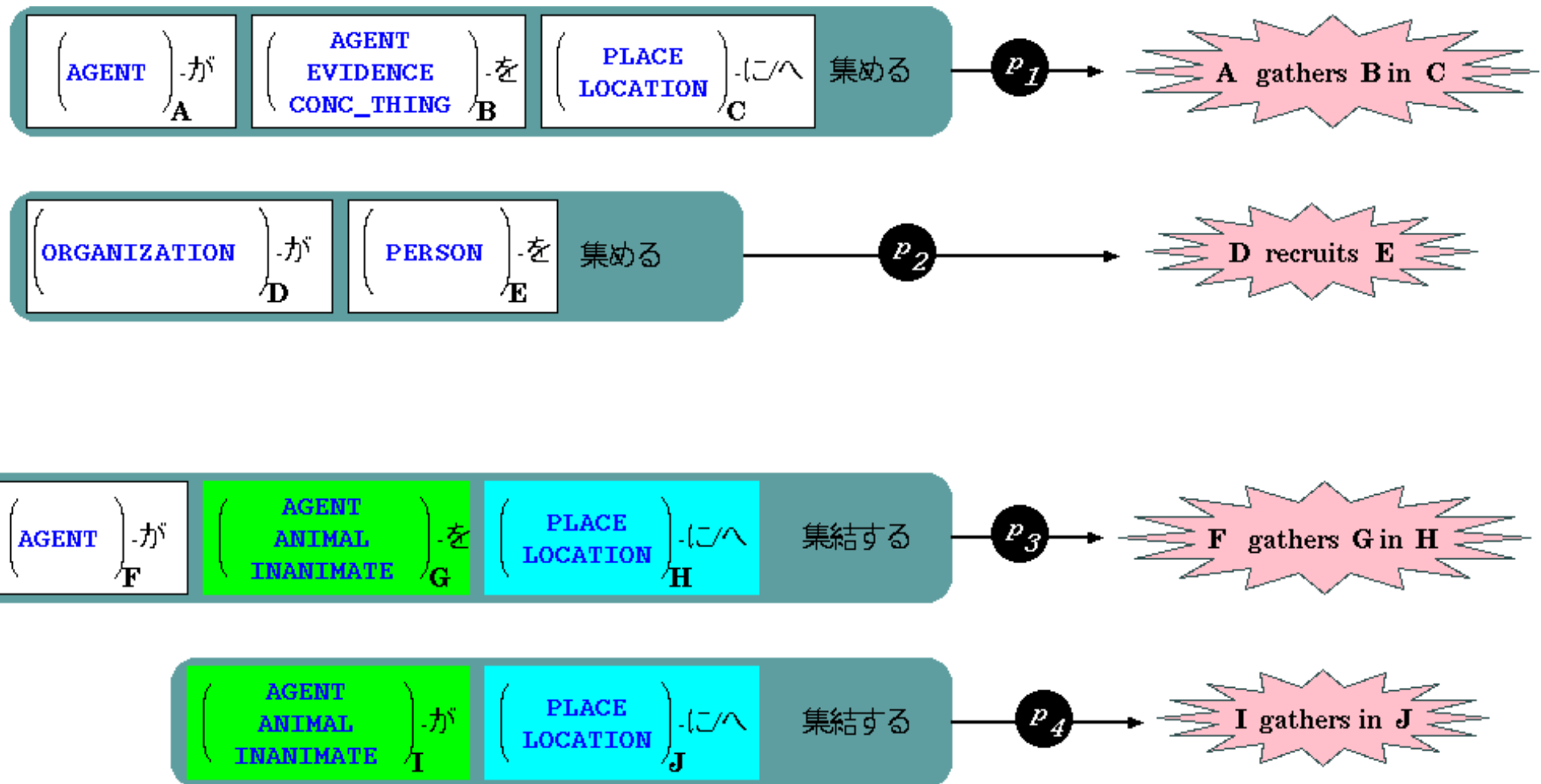
(Baldwin and Bond 2002)

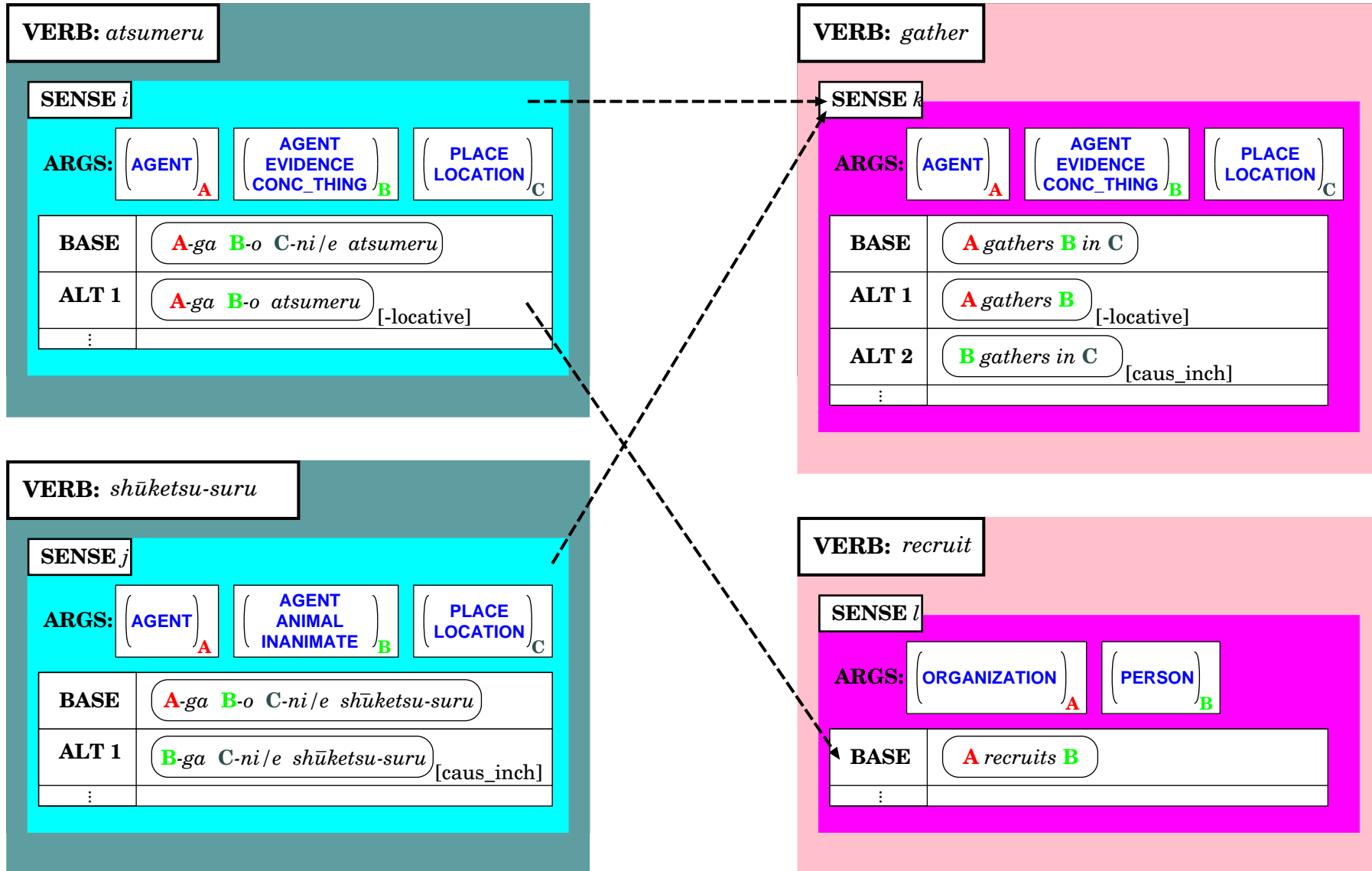
Basic Method

- Use selectional preferences to automatically extract alternations from a Japanese-English valency dictionary
- **Underlying hypothesis:** selectional preferences on alternating slots are the same
- Focus on Japanese verbs
- Analyse both the success of the method and what alternations we unearth

The Bigger Picture

- Move from a flat Japanese–English transfer dictionary to a hierarchical, language-modular dictionary structure
- In each monolingual lexicon, maximise structure sharing through analysis of alternations
- Assume no pre-defined alternation set (cf. Levin (1993)), no supervision in alternation extraction





Source Dictionary

- Goi-Taikei Japanese–English valency dictionary ◀
- Valency frame described in form of case frame headed by verb
- Each case slot annotated with:
 - ★ set of prototypical case markers
 - ★ POS (NP or S)
 - ★ set of selectional restrictions (→ Goi-Taikei thesaurus)
 - ★ set of lexical fillers

Constraints on Alternations

1. The selectional restrictions and lexical fillers on matching case slots are preserved under alternation
2. Alternations are monotonic in valency terms
3. A given alternation type has fixed direction: assume valency decreasing, and normalise direction alphabetically for valency-maintaining alternations (*over-constraint* ◀)

Extraction Procedure

1. Generate all legal alternation candidates for each case frame pairing (S, T) where S and T share some common kanji prefix
2. Score each, and return the highest scoring from among them
3. Accept only **non-negatively-scoring** alternations
4. In case of tie, select that alternation that preserves case marking the most

Scoring Alternations

- Score linked case slots S and T according to their relative *conceptual cohesion*:

$$cohesion(n_q) = -\log P(n_q) = -\log \frac{\sum_{lex_{p,i} \in n_q} freq(lex_{p,i})}{\sum_{lex_{p,i} \in n_o} freq(lex_{p,i})}$$

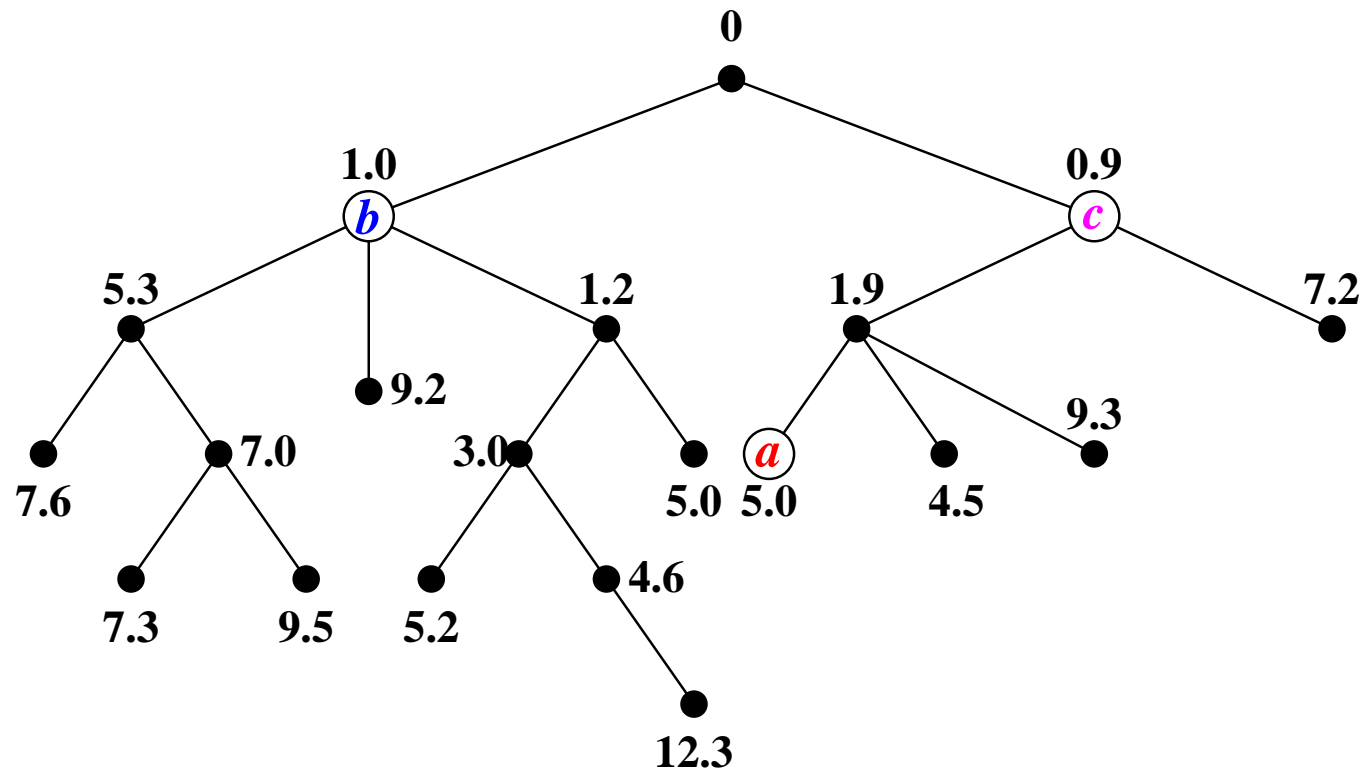
$$classmatch(n_j, n_k) = 3 cohesion(sub(n_j, n_k)) \\ - cohesion(n_j) - cohesion(n_k)$$

- Sum up the individual scores

$$\text{classmatch}(a, a) = 3 \times 5.0 - 5.0 - 5.0 = 5.0$$

$$\text{classmatch}(a, c) = 3 \times 0.9 - 0.9 - 5.0 = -3.2$$

$$\text{classmatch}(a, b) = 3 \times 0 - 1.0 - 5.0 = -6.0$$



Top 10 Extracted Alternations

<i>Index</i>	<i>Case slot mapping</i>		
1	$(NP_1\{ga\} \rightarrow \phi)$	$(NP_2\{o\} \rightarrow \{ga\})$	
2	$(NP_1\{ga\})$	$(NP_2\{o\} \rightarrow \phi)$	
3	$(NP_1\{ga\} \rightarrow \phi)$	$(NP_2\{o\} \rightarrow \{ga\})$	$(NP_3\{ni\})$
4	$(NP_1\{ga\} \rightarrow \phi)$	$(NP_2\{o\} \rightarrow \{ga\})$	$(NP_3\{ni, e\})$
5	$(NP_1\{ga\})$	$(NP_2\{o\} \rightarrow \phi)$	$(NP_3\{ni\} \rightarrow \{o\})$
6	$(NP_1\{ga\})$	$(NP_2\{o\})$	$(NP_3\{ni\} \rightarrow \phi)$
7	$(NP_1\{ga\})$	$(NP_2\{o\} \rightarrow \{kara, yori\})$	
8	$(NP_1\{ga\} \rightarrow \phi)$	$(NP_2\{o\} \rightarrow \{ga\})$	$(NP_3\{to, ni\})$
9	$(NP_1\{ga\})$	$(NP_2\{ni\} \rightarrow \{o\})$	
10	$(NP_1\{ga\} \rightarrow \phi)$	$(NP_2\{o\} \rightarrow \{ni\})$	$(NP_3\{de\} \rightarrow \{o\})$

Reflections

- Proposed method shown to be effective in extracting out valid alternations
- Little sense of recall (although not necessarily important for the dictionary reconstruction process)
- Possibility for using translation information to improve the accuracy of the extraction method

A General Feature Space for Automatic Verb Classification

(Joanis and Stevenson 2003)

Basic Method

- Use alternations and general verbal features to classify verbs according to Levin (1993) classes
- Dodge the issue of alternation detection or subcat acquisition by relying on features which capture alternation effects only indirectly
- Supplement alternation-based features with various weak lexical semantic indicators

Syntactic Slot-based Features

- Frequency of different syntactic slots occurring with a verb (includes PPs, conditioned on P)
- Degree of lexical overlap between syntactic slots known to alternate
- Expletive pronouns/*there*

Tense, Voice and Aspect Features

- Relative frequency of passivisation
- POS (tense) of the verb
- Relative occurrence with modals/adverbials
- Relative occurrence in derived forms

Animacy Feature

- Relative occurrence of animate fillers (personal pronouns, person names) in each of the syntactic slots

Task

- 2/3-way classification of a range of verb classes:
 - ★ benefactive vs. recipient verbs
 - ★ *admire* vs. *amuse* verbs
 - ★ *run* vs. sound emission verbs
 - ★ *cheat* vs. *cheat/steal* verbs
 - ★ *wipe* vs. *cheat/steal* verbs
 - ★ *spray/load* vs. *fill* vs. other *put* verbs
 - ★ *run* vs. change of state vs. object drop verbs
- Also combined multi-way tasks

Experiments

- Feature values extracted from BNC (parsed with SCOL)
- Focus on verbs which occur > 100 times in the BNC in only one of the classes under consideration (with the predominant sense), and which are not excessively polysemous
- C5.0 used as learner (decision tree-based)
- Varied results were obtained

Reflections

- General technique proposed for verbal classification, based partly on alternation behaviour
- Little sense of what works well for what class, or, e.g., whether selectional preferences aid the classifier
- Potential for improvement through subcat frame acquisition (remove independence of syntactic slots), explicit modelling of selectional preferences and a better parser

Decision Tree Learning

Constructing Decision Trees: ID3

- **Basic method:** construct decision trees in recursive divide-and-conquer fashion

FUNCTION ID3 (Root)

IF all instances at root have same class

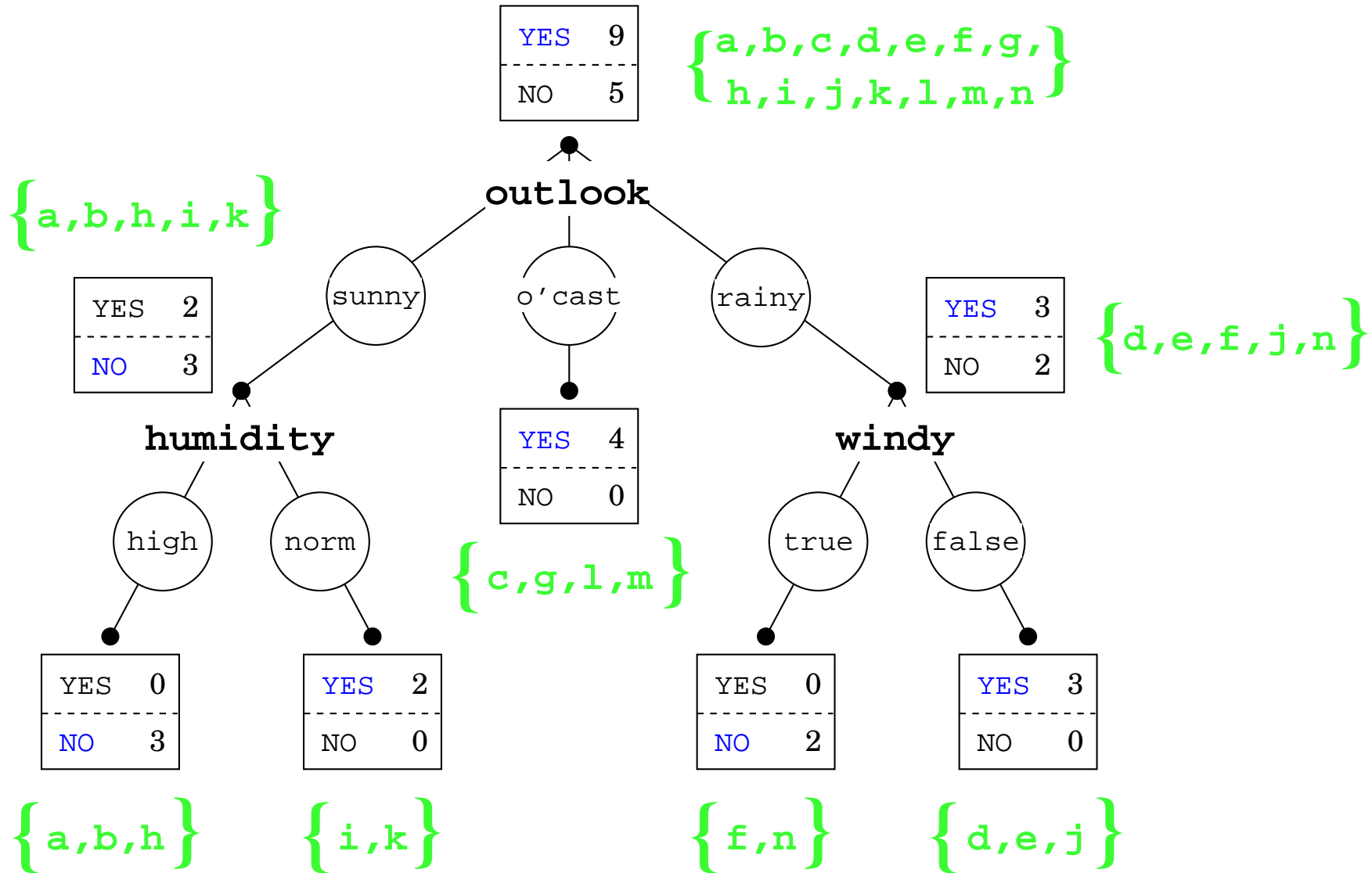
THEN stop

ELSE Select a new attribute to use in partitioning root node instances

Create a branch for each attribute value and partition up root node instances according to each value

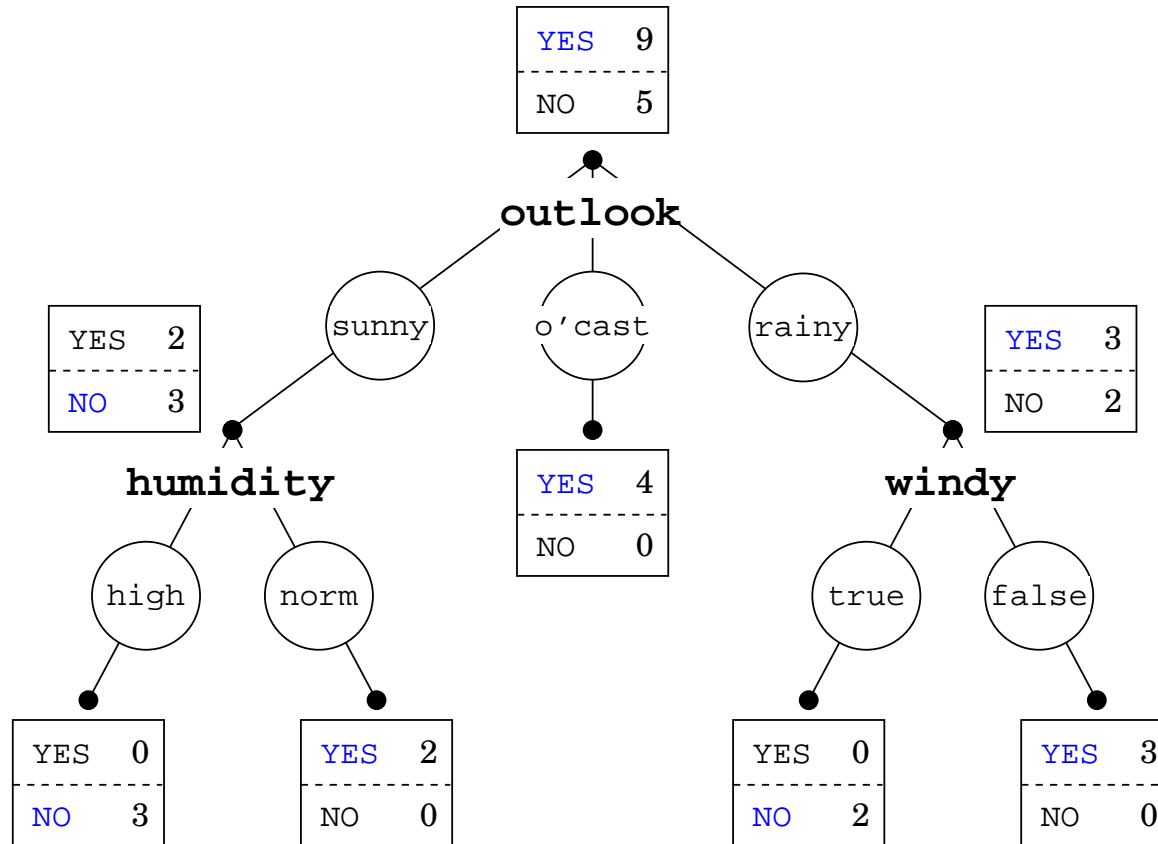
Call ID3(LEAF_{*i*}) for each leaf node LEAF_{*i*}

- Note: we may end up with non-pure leaves



Classifying Novel Instances

- Having constructed the decision tree, we classify novel instances by traversing down the tree and classifying according to the majority class at the deepest reachable point in the tree structure
- Complications:
 - ★ unobserved attribute–value pairs
 - ★ missing values



TEST DATA

(sunny, hot, normal, FALSE)
 (rainy, hot, low, FALSE)
 (?, cool, high, TRUE)

Criterion for Attribute Selection

- Which is the best attribute?
 - ★ want to get the smallest tree (Occam's Razor; generalisability)
- **Heuristic:** choose the attribute that produces the “purest” nodes according to **information gain** (IG)
 - information gain increases with the average purity of the subsets
- **Strategy:** choose the attribute that gives the greatest information gain
- NB standard vs. oblivious decision trees

Mean Information Associated with a Decision Stump

- We calculate the mean information for a tree stump with m attributes as:

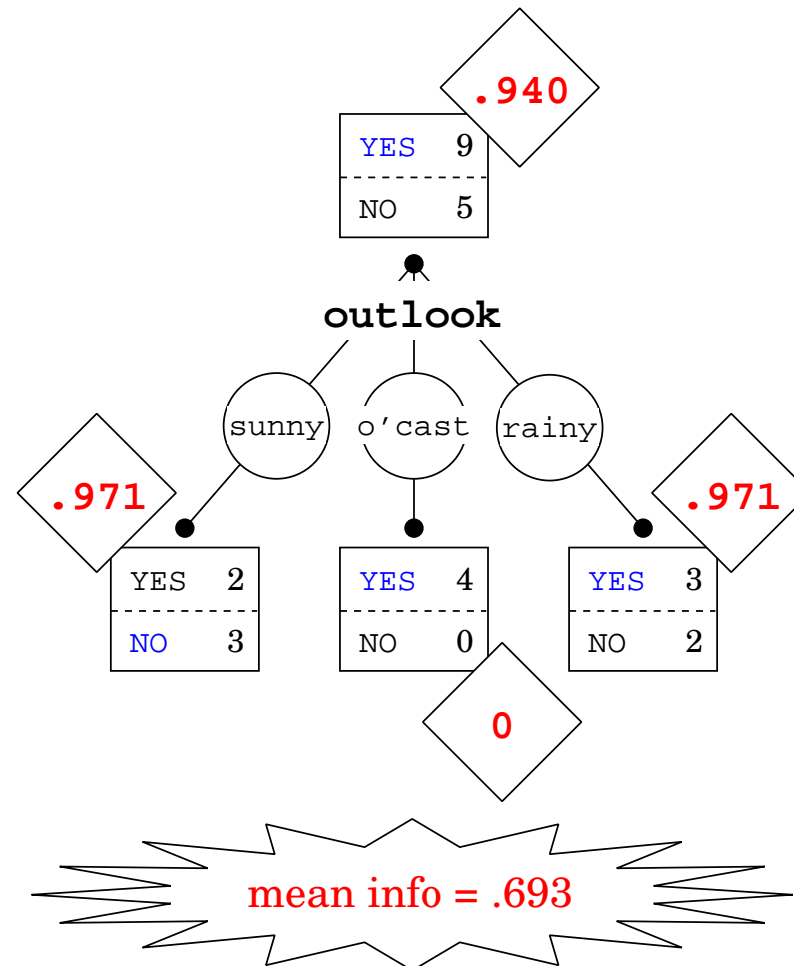
$$H(x_1, \dots, x_m) = \sum_{i=1}^m P(x_i) H(x_i)$$

where $H(x_i)$ is the entropy at node x_i

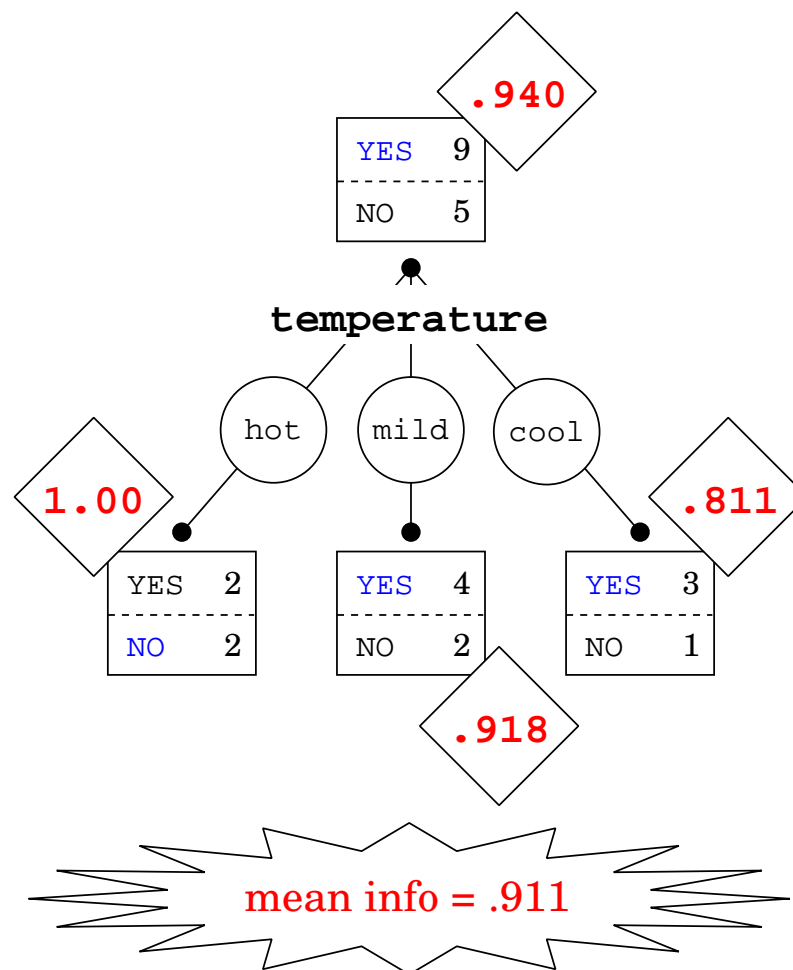
Full weather.nominal Dataset

	Outlook	Temperature	Humidity	Windy	Play
a:	sunny	hot	high	FALSE	no
b:	sunny	hot	high	TRUE	no
c:	overcast	hot	high	FALSE	yes
d:	rainy	mild	high	FALSE	yes
e:	rainy	cool	normal	FALSE	yes
f:	rainy	cool	normal	TRUE	no
g:	overcast	cool	normal	TRUE	yes
h:	sunny	mild	high	FALSE	no
i:	sunny	cool	normal	FALSE	yes
j:	rainy	mild	normal	FALSE	yes
k:	sunny	mild	normal	TRUE	yes
l:	overcast	mild	high	TRUE	yes
m:	overcast	hot	normal	FALSE	yes
n:	rainy	mild	high	TRUE	no

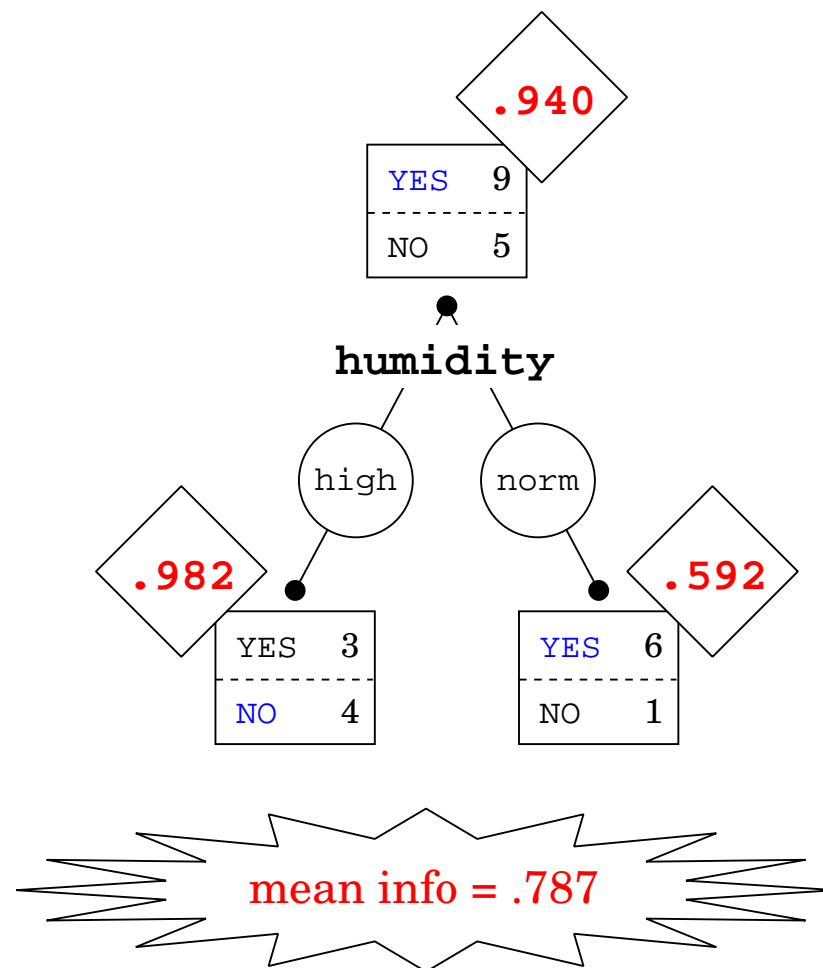
Mean Information (outlook)



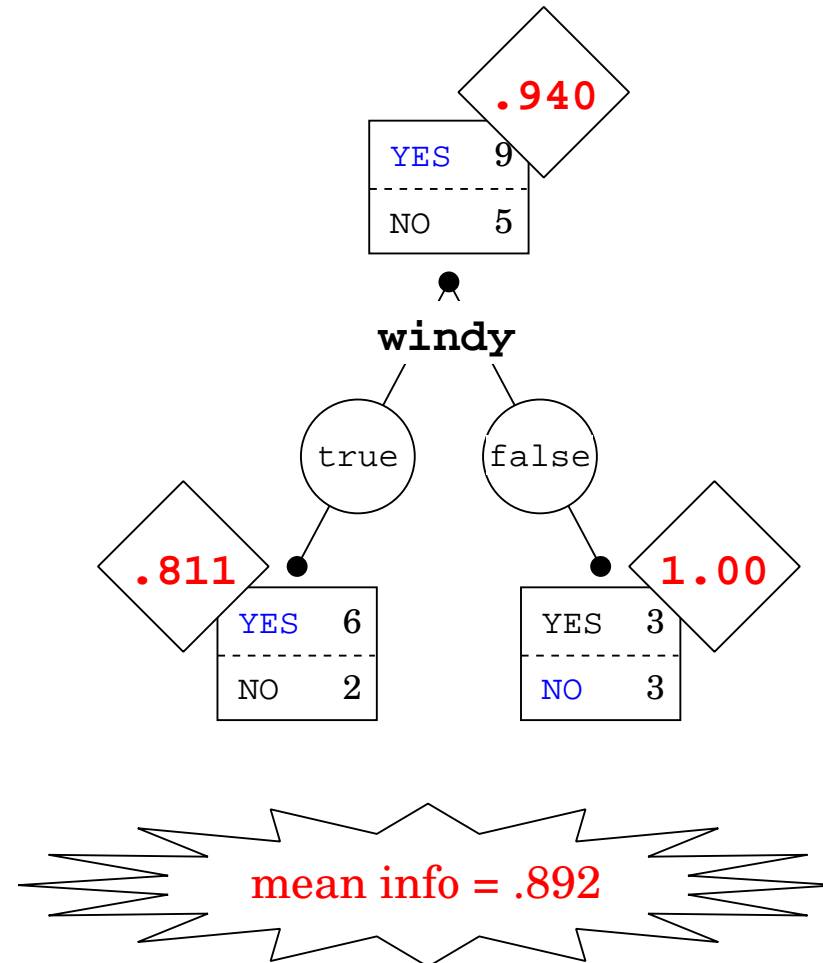
Mean Information (temperature)



Mean Information (humidity)



Mean Information (windy)



Attribute Selection: Information Gain

- We determine which attribute R_A (with values x_1, \dots, x_m) best partitions the instances at a given root node R according to information gain:

$$IG(R_A|R) = H(R) - \sum_{i=1}^m P(x_i)H(x_i)$$

$$IG(\text{outlook}|R) = 0.247$$

$$IG(\text{temperature}|R) = 0.029$$

$$IG(\text{humidity}|R) = 0.152$$

$$IG(\text{windy}|R) = 0.048$$

References

- BALDWIN, TIMOTHY, and FRANCIS BOND. 2002. Alternation-based lexicon reconstruction. In *Proc. of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2002)*, 1–11, Keihanna, Japan.
- BRENT, MICHAEL R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics* 19.243–62.
- JOANIS, ERIC, and SUZANNE STEVENSON. 2003. A general feature space for automatic verb classification. In *Proc. of the 10th Conference of the EACL (EACL 2003)*, 163–70, Budapest, Hungary.
- LEVIN, BETH. 1993. *English Verb Classes and Alterations*. Chicago, USA: University of Chicago Press.
- MANNING, CHRISTOPHER D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proc. of the 31st Annual Meeting of the ACL*, 235–42.