

ARGUMENTATION MINING

Marie-Francine Moens joint work with Raquel
Mochales Palau and Parisa Kordjamshidi
Language Intelligence and Information
Retrieval
Department of Computer Science
KU Leuven, Belgium
Dundee, 5-9-2014

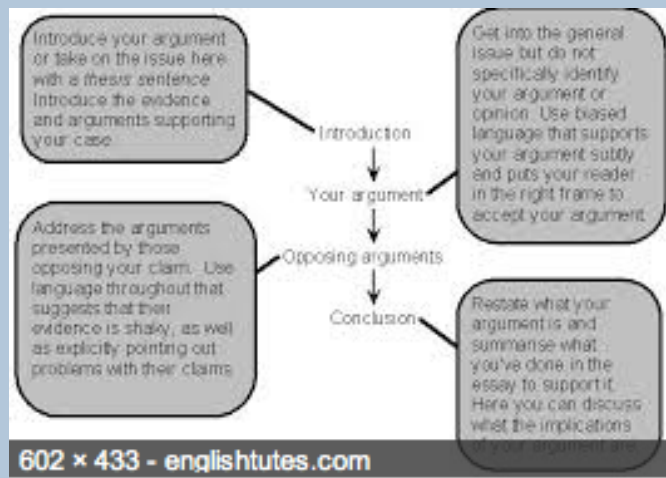
OUTLINE

- **Part 1: The setting**
 - Definition of argumentation mining
 - Importance of the task
- **Part 2: Introducing current methods**
 - Machine learning
 - Features
 - Common techniques: logistic regression, conditional random fields, support vector machines
 - Joint recognitions: grammars, graphical models, structured support vector machines
 - Features revisited
 - Textual entailment
- **Part 3: Some applications**
 - Legal field
 - Scientific texts
 - Blogs
 - Dialogues and debates
- **Part 4: Conclusions and thoughts for future research**

■ PART 1: The setting

ARGUMENTATION MINING

- = the detection of an argumentative discourse structure in text or speech, and the detection and the functional classification of its composing components



ARGUMENTATION MINING

- **Argumentation mining = recognition of a rhetorical structure in a discourse**
- **Rhetoric is the art of discourse that aims to improve the capabilities of writers and speakers to inform, persuade or motivate particular audiences in specific situations**

[Corbett, E. P. J. (1990). *Classical rhetoric for the modern student*. New York: Oxford University Press., p. 1.]

ARGUMENTATION

- Is probably as old as mankind
- Has been studied by philosophers throughout the history



Painting depicting a lecture in a knight academy, painted by [Pieter Isaacsz](#) or [Reinhold Timm](#) for [Rosenborg Castle](#) as part of a series of seven paintings depicting the seven independent arts. This painting illustrates rhetorics



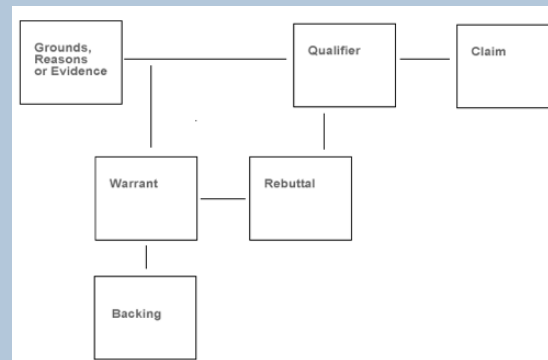
SOME HISTORY

- From Ancient Greece to the late 19th century: central part of Western education: need to train public speakers and writers to move audiences to action with arguments
- The approach of argumentation is very often based on theories of rhetoric and logic
- Argumentation was/is taught at universities

SOME HISTORY

■ Highlights:

- Aristotle's (4th century BC) logical works: *Organon*
- George Pierce Baker, *The Principles of Argumentation*, 1895
- Chaïm Perelman describes of techniques of argumentation used by people to obtain the approval of others for their opinions: *Traité de l'argumentation – la nouvelle rhétorique*, 1958
- Stephen Toulmin explains how argumentation occurs in the natural process of an everyday argument: *The Uses of Argument*, Cambridge University Press, 1958



Argumentation in text

One of most the fundamental things we use language for is argument. Arguing means claiming that something is true and trying to persuade other people to agree with your claim by presenting evidence to substantiate it. An argument is statement with three components:

1. A point of view, a claim, something we are arguing *in favour of*
2. The actual argument, the evidence we are using to argue *with*
3. A statement that links the initial claim to the argument and ensures that we understand how the argument functions.

The statement that connects the initial claim and the argument is referred to as the warrant. The warrant is thus an argument for the connection between the initial claim and the argument.

<http://sokogskriv.no/en/reading/argumentation-in-text/>

TODAY

■ We find argumentation in:

- Legal texts and court decisions
- Biomedical cases
- Scientific texts
- Patents
- Reviews, online fora, user generated content
- Debates, interactions, dialogues
- ...

Claims:

1. A method for determining a probabilistic, context dependent word distribution for each word in a previously unseen text, the method comprising: in a training phase, learning for each word of a large corpus of natural language texts a probabilistic context model that describes the context these words typically occur in and learning a hidden-to-observed distribution that describes words with similar meaning and usage; storing the context model and the hidden-to-observed distribution on a storage device; and in an inference phase, retrieving the context model and the hidden-to-observed distribution from the storage device and for each word in the previously unseen text determining the probabilistic, context dependent word distribution utilizing the context model and the hidden-to-observed distribution obtained in the training phase.

2. The method according to claim 1 wherein, in the training phase, the probabilistic context model and the context dependent word distribution are iteratively refined.

3. The method according to claim 1 wherein the training phase comprises: tokenizing the corpus of natural language texts into individual words; representing the corpus of natural language text with a Bayesian model with a hidden or latent variable for every word in the corpus, the Bayesian model representing the context dependent set of similar words, and with dependencies between the hidden variable and the hidden variables in its context, the dependencies representing the context model, and with dependencies between the hidden variable and the observed word at that position, the dependencies representing the hidden-to-observed distribution; and utilizing approximate inference methods to determine a probabilistic distribution of words for the hidden variables, to learn the context model and to learn the hidden-to-observed distribution.

4. The method according to claim 2 wherein the training phase comprises: tokenizing the corpus of natural language texts into individual words; representing the corpus of natural language text with a Bayesian model with a hidden or latent variable for every word in the corpus, the Bayesian model representing the context dependent set of similar words, and with dependencies between the hidden variable and the hidden variables in its context, the dependencies representing the context model, and with dependencies between the hidden variable and the observed word at that position, the dependencies representing the hidden-to-observed distribution; and utilizing approximate inference methods to determine a probabilistic distribution of words for the hidden variables, to learn the context model and to learn the hidden-to-observed distribution.

5. The method according to claim 1 wherein the inference phase comprises: tokenizing the text into individual words; representing the text with a Bayesian model with a hidden or hidden variable for every word in the corpus, the Bayesian model representing the context dependent set of similar words, and with dependencies between the hidden variable and the hidden variables in its context and between the hidden variable and the observed word at that position; and utilizing the context model and the hidden-to-observed distribution learned in the training phase together with approximate inference methods to determine a probabilistic distribution of words for the hidden variables in a previously unseen text.

6. The method according to claim 2 wherein the inference phase comprises: tokenizing the text into individual words; representing the text with a Bayesian model with a hidden or hidden variable for every word in the corpus, the Bayesian model representing the context dependent set of similar words, and with dependencies between the hidden variable and the hidden variables in its context and between the hidden variable and the observed word at that position; and utilizing the context model and the hidden-to-observed distribution learned in the training phase together with approximate inference methods to determine a probabilistic distribution of words for the hidden variables in a previously unseen text.

7. The method according to claim 3 wherein the inference phase comprises: tokenizing the text into individual words; representing the text with a Bayesian model with a hidden or hidden variable for every word in the corpus, the Bayesian model representing the context dependent set of similar words, and with dependencies between the hidden variable and the hidden variables in its context and between the hidden variable and the observed word at that position; and utilizing the context model and the hidden-to-observed distribution learned in the training phase together with approximate inference methods to determine a probabilistic distribution of words for the hidden variables in a previously unseen text.

8. The method according to claim 4 wherein the inference phase comprises: tokenizing the text into individual words; representing the text with a Bayesian model with a hidden or hidden variable for every word in the corpus, the Bayesian model representing the context dependent set of similar words, and with dependencies between the hidden variable and the hidden variables in its context and between the hidden variable and the observed word at that position; and utilizing the context model and the hidden-to-observed distribution learned in the training phase together with approximate inference methods to determine a probabilistic distribution of words for the hidden variables in a previously unseen text.

9. A method for automatic analysis of natural language, the method comprising: utilizing a probabilistic, context dependent word distribution determined by the method according to claim 1 for each word in a previously unseen text.



1920 x 1200 - technmarketing.com

WHY ARGUMENTATION MINING?

- In the overload of information users want to find arguments that sustain a certain claim or conclusion
- Argumentation mining refines:
 - Search and information retrieval
 - Provides the end user with instructive visualizations and summaries of an argumentative structure

Argumentation mining is related to opinion mining, but end user wants to know the underlying grounds and maybe counterarguments

WHAT IS THE STATE-OF-THE-ART?

- Argumentative zoning
- Argumentation mining of legal cases
- Argumentation mining in online user comments and discussions
- ...

ARGUMENTATIVE ZONING

- = segmentation of a discourse into discourse segments or zones that each play a specific rhetoric role in a text

Distributional Clustering of English Words
Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Unsupervised annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuition in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Problem Setting

In what follows, we will consider two major word classes, <BN> and <BN>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <BN> of occurrence of particular pairs <BN> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser: Piddich (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yanowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, so what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs, the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any many relation in terms of associations between elements in each coordinate and appropriate hidden units (clusters, centroids) and associations between these hidden units.

BKG: General scientific background (yellow)

OTH: Neutral descriptions of other people's work (orange)

OWN: Neutral descriptions of the own, new work (blue)

AIM: Statements of the particular aim of the current paper (pink)

TXT: Statements of textual organization of the current paper (in chapter 1, we introduce...)
(red)

CTR: Contrastive or comparative statements about other work; explicit mention of weaknesses of other work (green)

BAS: Statements that own work is based on other work (purple)

[PHD thesis of Simone Teufel 2000]

ARGUMENTATIVE ZONING

- Methods: seen as a classification task: rule based or statistical classifier (e.g., naïve Bayes, support vector machine) is trained with manually annotated examples

[Moens, M.-F. & Uyttendaele, C. *Information Processing & Management* 1997]

[Teufel, S. & Moens, M. *ACL* 1999]

[Teufel, S. & Moens, M. *EMNLP* 2000]

[Hachey, B. & Grover, C. *ICAIL* 2005]

ARGUMENTATION MINING OF LEGAL CASES

- Legal field:
 - Precedent reasoning
 - Search for cases that use a similar type of reasoning, e.g., acceptance of rejection of a claim based on precedent cases
 - Adds an additional dimension to argumentative zoning:
 - Needs detection of the argumentation structure and classification of its components
 - Components or segments are connected with argumentative relationships

[Moens, Boiy, Mochales & Reed ICAIL 2007]

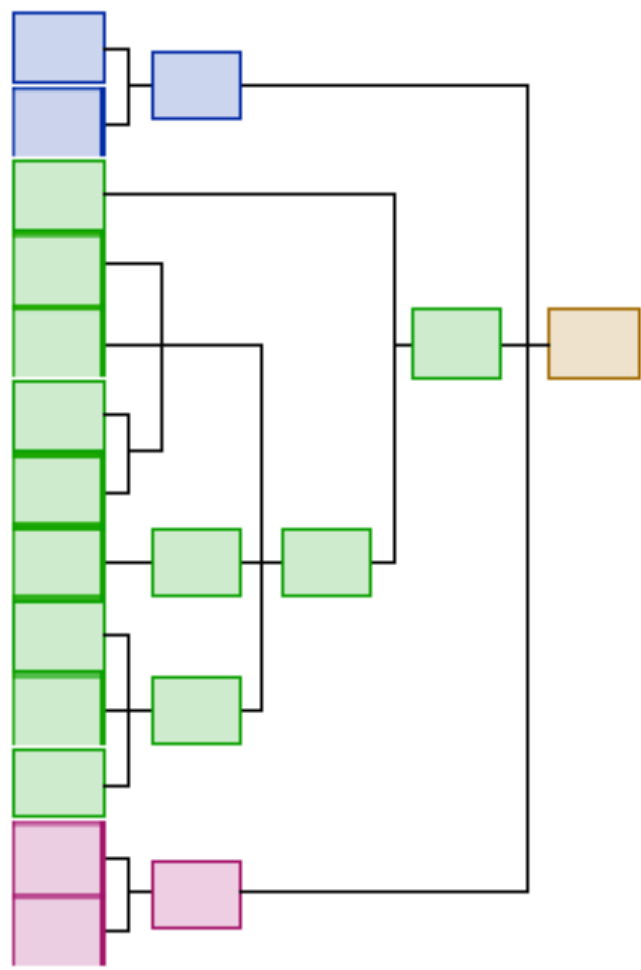


Figure 1.1: Reasoning structure of the legal case in Appendix A. Each block is a sentence of the legal case. There are 3 arguments (blue, green and red) that justify the final decision (brown). The contents of each argument and the final decision can be seen in detail in Figures 1.2, 1.3, 1.4 and 1.5

[PhD thesis Raquel Mochales Palau 2011]

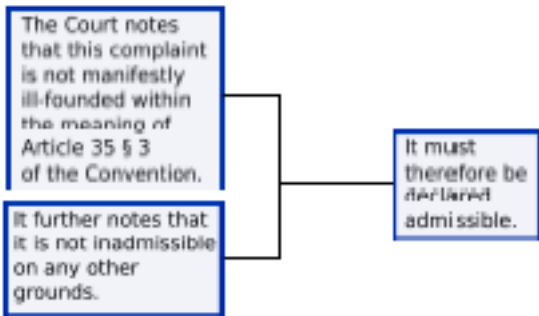


Figure 1.2: Closer view 1st Argument

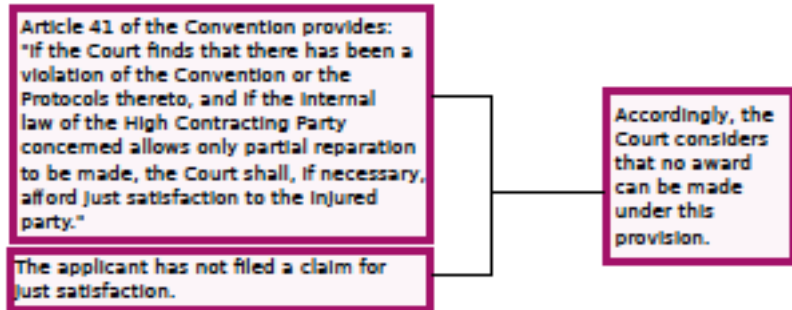


Figure 1.3: Closer view 2nd Argument

[PhD thesis Raquel Mochales Palau 2011]

FOR THESE REASONS, THE COURT UNANIMOUSLY
 1. Declares the application admissible;
 2. Holds that there has been a violation of Article 6 § 1 of the Convention

Figure 1.4: Closer view Final Decision

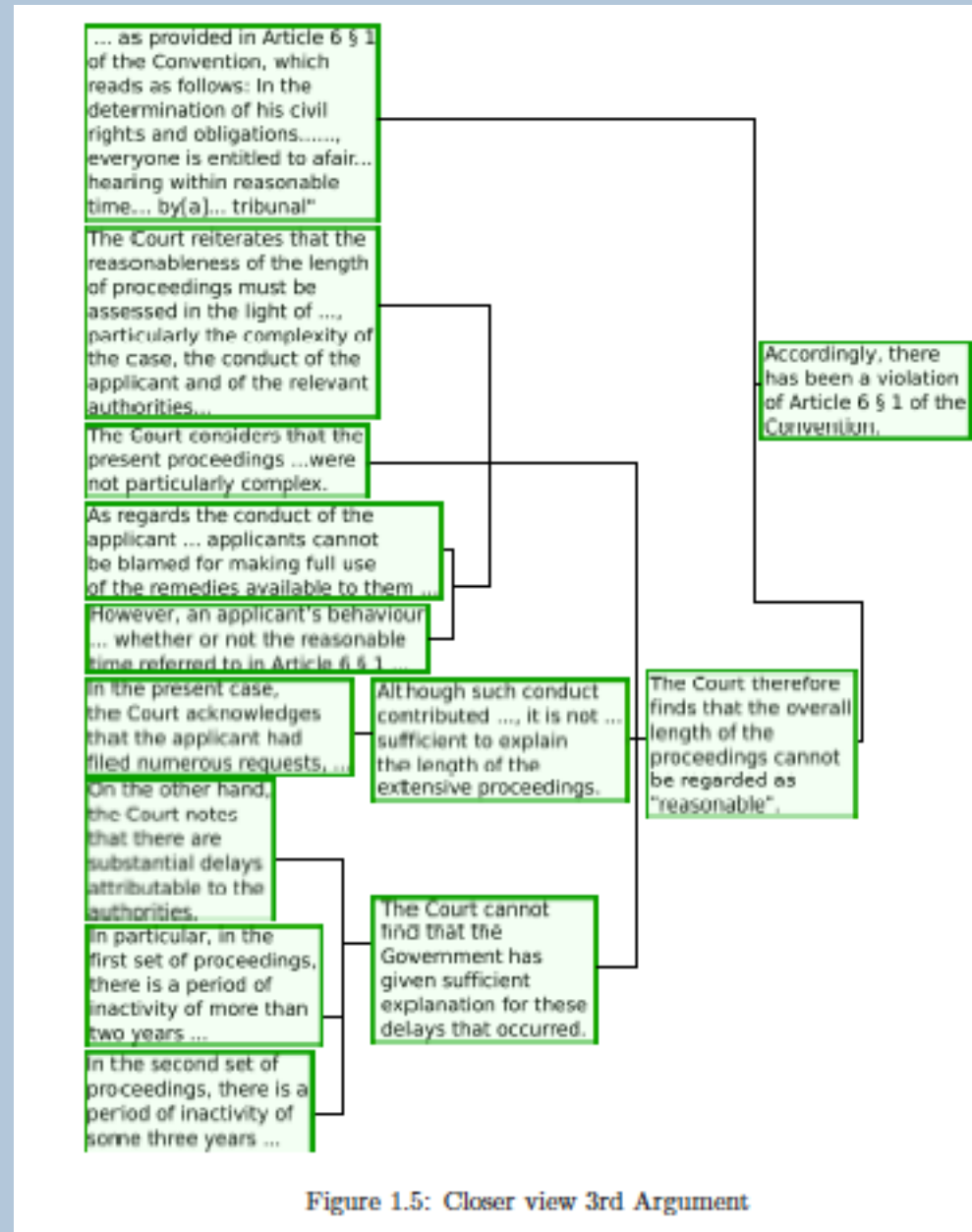


Figure 1.5: Closer view 3rd Argument

- [PhD thesis of Raquel Mochales 2011]

- Argumentation: a process whereby arguments are constructed, exchanged and evaluated in light of their interactions with other arguments
- **Argument**: a set of **premises** - pieces of evidence - in support of a **claim**
- **Claim**: a proposition, put forward by somebody as true; the claim of an argument is normally called its conclusion
- Argumentation may also involve chains of reasoning, where claims are used as premises for deriving further claims

```

T
|--D
| |--: For these reasons, the Commission by a majority declares the application admissible,
|       without prejudging the merits.
|--A
|   |--A
|   | |--C
|   | |--: It follows that the application cannot be dismissed as manifestly ill-founded.
|   | |--A
|   |   |--P
|   |   |--: It considers that the applicant 's complaints raise serious issues of fact
|   |   |       and law under the convention, the determination of which should depend on
|   |   |       an examination of the merits.
|   |   |--P
|   |   |--: The Commission has taken cognizance of the submissions of the parties.
|--A
|   |--C
|   |--: In these circumstances, the Commission finds that the application cannot be
|       declared inadmissible for non-exhaustion of domestic remedies.
|--A
|   |--P
|   |--: The Commission recalls that article art. x of the convention only requires
|       the exhaustion of such remedies which relate to the breaches of the
|       convention alleged and at the same time can provide effective and sufficient
|       redress.
|   |--P
|   |--: The Commission notes that in the context of the section powers the
|       secretary of state has a very wide discretion.
|--P
|   |--: The Commission recalls that in the case of temple v. the united kingdom
|       no. x dec. d.r. p.
|--P
|   |--: The Commission held that recourse to a purely discretionary power on
|       the part of the secretary of state did not constitute an effective
|       domestic remedy.
|   |--: The Commission finds that the suggested application for discretionary
|       relief in the instant case cannot do so either.

```

Fig. 6: Output of the automatic system: small fragment of the argumentation tree-structure of a document

■ Part 2: Introducing current methods

TEXT MINING

Text mining, also referred to as *text data mining*, or roughly *equivalent to text analytics*:

= deriving high quality information from text

Often done through means of statistical patterns learning

⇒ Use of statistical machine learning techniques

ARGUMENTATION MINING

- Because argumentation is well studied: typical argumentation structures are defined:
- => structuring of information: detecting the argumentation and its components
- => assignment of metadata: labeling of argumentation components and relations
- Can be done manually:
 - But, people are (often) expensive, slow and inconsistent
 - Can we perform this task automatically?

ARGUMENTATION MINING

- Approaches: pattern recognition
 - **Symbolic techniques:** knowledge or part of it:
 - formally and manually implemented
 - **Statistical machine learning techniques:** knowledge or part of it:
 - automatically acquired

ARGUMENTATION MINING

- **Mostly: using supervised machine learning techniques**
- **Why ?**
 - Argumentation structure is well studied
 - Manually labeled examples are available
 - Annotating examples is usually considered easier than pattern engineering
 - Current supervised learning techniques allow integration of soft rules

ARGUMENTATION MINING

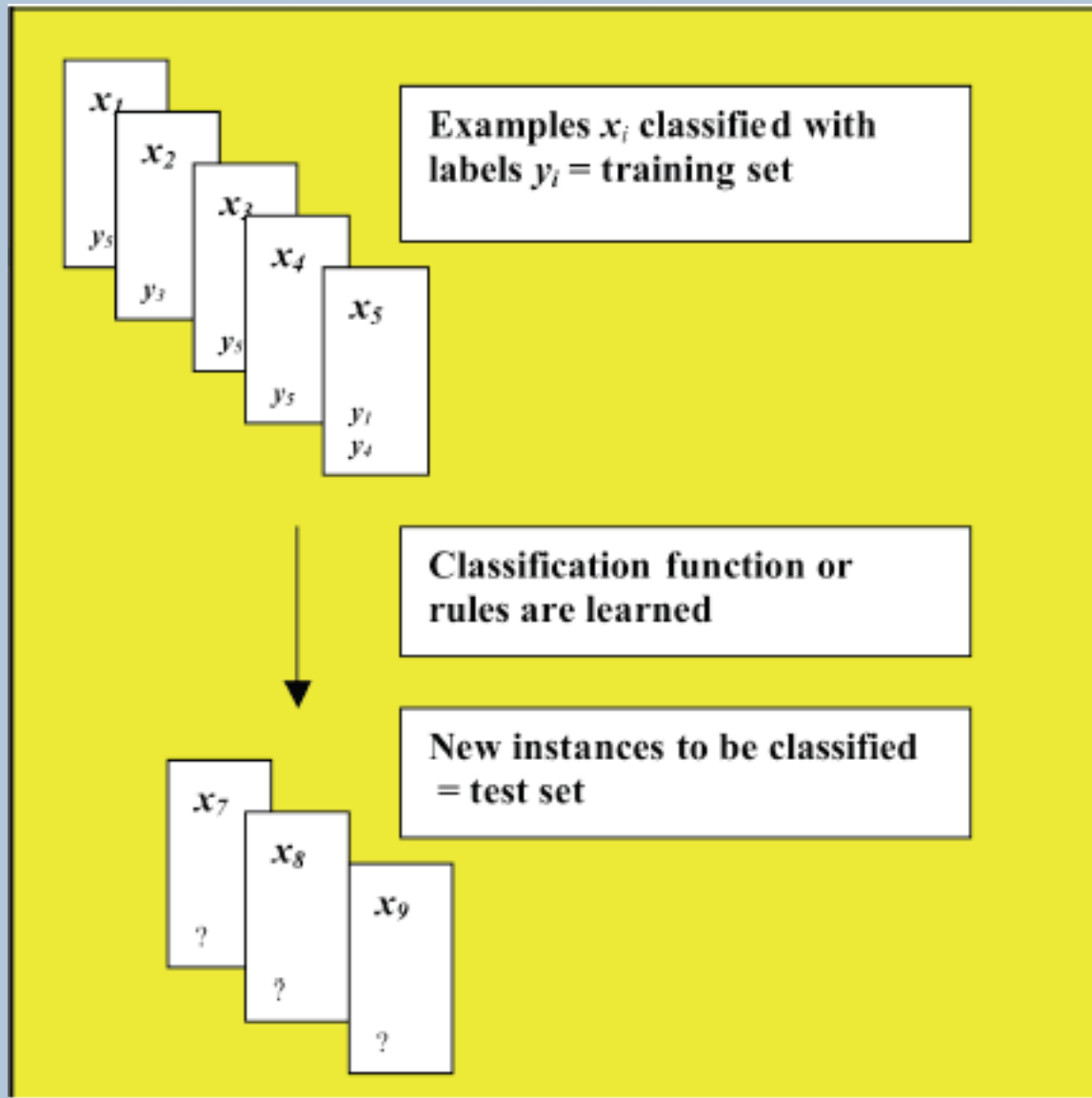
- Argumentation mining needs a large amount of knowledge:
 - Linguistic knowledge of the vocabulary, syntax and semantics of the language and the discourse
 - Knowledge of the subject domains
 - Background knowledge of the person who uses the texts at a certain moment in time

SUPERVISED LEARNING

- **Techniques of supervised learning:**
 - **training set:** example objects classified by an expert or teacher
 - detection of general, but high-accuracy classification patterns (function or rules) in the training set based on object features and their values
 - patterns are predictable to correctly classify new, previously unseen objects in a **test set** considering their features and feature values

SUPERVISED LEARNING

- Text recognition or classification can be seen as a:
 - two-class learning problem:
 - an object is classified as belonging or not belonging to a particular class
 - convenient when the classes are not mutually exclusive
 - single multi-class learning problem
- Result = often **probability** of belonging to a class, rather than simply a classification



GENERATIVE VERSUS DISCRIMINATIVE CLASSIFICATION

- In classification: given inputs \mathbf{x} and their labels y :
 - **Generative classifier** learns a model of the joint probability $p(\mathbf{x}, y) = p(y) p(\mathbf{x}|y)$ and then condition on the observed features \mathbf{x} , thereby deriving the class posterior $p(y|\mathbf{x})$ and selects the most probable y for \mathbf{x}
 - Generative classifier: since it specifies how to generate the observed features \mathbf{x} for each class y
 - E.g.,
 - Naive Bayes, hidden Markov model

GENERATIVE VERSUS DISCRIMINATIVE CLASSIFICATION

- **Discriminative classifier learns a model** $p(y|x)$ which directly models the mapping from inputs x to output y , and selects the most likely label y
- Discriminative classifier: discriminates between classes
- E.g.,
 - Logistic regression model, conditional random field, support vector machine

(discussed in this tutorial)

MAXIMUM ENTROPY PRINCIPLE

- Text classifiers are often trained with **incomplete information**
- Probabilistic classification can adhere to the principle of maximum entropy: When we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we have: e.g.,
 - Multinomial logistic regression, conditional random fields

CONTEXT-DEPENDENT RECOGNITION

- When there exist a relation between various classes: it is valuable not to classify an object separately from other objects
- **Context-dependent classification:** the class to which a feature vector is assigned depends on:
 - the object itself
 - other objects and their class
 - the existing relations among the various classes
 - e.g., hidden Markov model, conditional random fields, structured support vector machine, structured perceptron

LOCAL VERSUS GLOBAL CLASSIFICATION

- Local classification (i.e., learning a model for each class), applying the models on each input, and combining the outputs
- **Global classification** (i.e., learning 1 model jointly, cf. context dependent classification)

FEATURE SELECTION AND EXTRACTION

- In classification tasks: object is described with set of **attributes or features**
- Typical features in text classification tasks:
 - word, phrase, syntactic class of a word, text position, the length of a sentence, the relationship between two sentences, an n -gram, a document (term classification),
 - choice of the features is application- and domain-specific
- Features can have a value, for text the value is often:
 - numeric, e.g., discrete or real values
 - nominal, e.g. certain strings
 - ordinal, e.g., the values 0= small number, 1 = medium number, 2 = large number

FEATURE SELECTION AND EXTRACTION

- The features together span a multi-variate space called the measurement space or feature space:
 - an object x can be represented as:
 - a **vector of features**:
$$x = [x_1, x_2, \dots, x_p]^T$$
where p = the number of features measured
 - as a **structure**: e.g.,
 - representation in first order predicate logic
 - graph representation (e.g., tree) where relations between features are figured as edges between nodes and nodes can contain attributes of features

Examples of classification features

SWARM INTELLIGENCE

Following a trail of insects as they work together to accomplish a task offers unique possibilities for problem solving.

By Peter Tarasewich & Patrick R. McMullen

Even with today's ever-increasing computing power, there are still many types of problems that are very difficult to solve. Particularly combinatorial optimization problems continue to pose challenges. An example of this type of problem can be found in product design. Take as an example the design of an automobile based on the attributes of engine horsepower, passenger seating, body style and wheel size. If we have three different levels for each of these attributes, there are 3^4 , or 81, possible configurations to consider. For a slightly larger problem with 5 attributes of 4 levels, there are suddenly 1,024 combinations. Typically, an enormous amount of possible combinations exist, even for relatively small problems. Finding the optimal solution to these problems is usually impractical. Fortunately, search heuristics have been developed to find good solutions to these problems in a reasonable amount of time.

Over the past decade or so, several heuristic techniques have been developed that build upon observations of processes in the physical and biological sciences. Examples of these techniques include Genetic Algorithms (GA) and simulated annealing...

Sentence position

Words

POS-tag

The following and preceding word

Sentence length

FEATURE VECTORS FOR AN EXAMPLE TEXT

A Java Applet that scans Java Applets

- Binary values, based on lower-cased words:
[a: 1, apple: 0, applet: 1, applets: 1, ..., java: 1, ...]
- Remove stopwords :
[apple : 0, applet : 1, applets : 1, ... , java : 1 ...]
- Numeric value: based on text term frequency (*tf*):
[apple : 0, applet : 1, applets : 1, ... , java : 2 ...]
- Numeric value: based on text term frequency of lower cased *n-grams* (*tf*):
[aa: 0, a_a: 2, a_b: 0, ...]
- Numeric attribute value based on latent semantic indexing:
[F1: 0.38228938, F2: 0.000388, F3: 0.201033, ...]
- ...

FEATURE SELECTION

- = eliminating low quality features:
 - redundant features
 - noisy features
- Decreases computational complexity
- Decreases the danger of overfitting in supervised learning (especially when large number of features and few training examples)
- **Overfitting:**
 - the classifier perfectly fits the training data, but fails to generalize sufficiently from the training data to correctly classify the new case

FEATURE EXTRACTION

- = creates new features by applying a set of operators upon the current features:
 - a single feature can be replaced by a new feature (e.g., replacing words by their stem)
 - a set of features is replaced by one feature or another set of features
 - use of logical operators (e.g., disjunction), arithmetical operators (e.g. mean, LSI)
 - choice of operators: application- and domain-specific

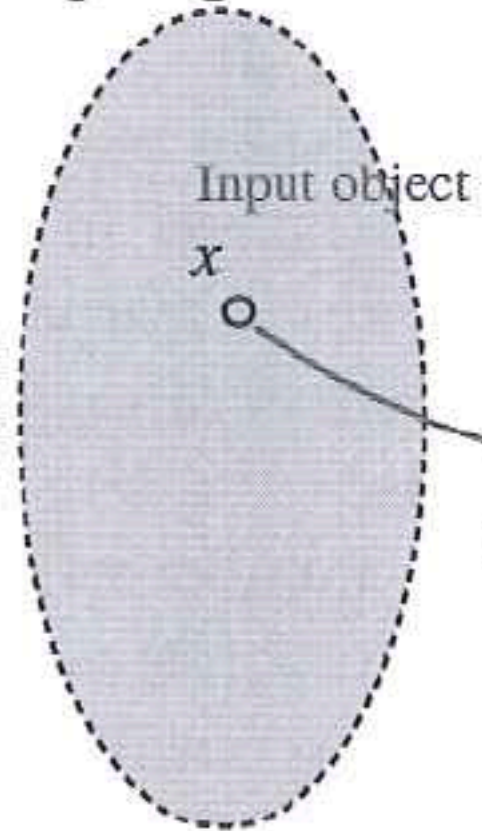
COMMON CLASSIFICATION METHODS

- Naïve Bayes, learning of rules and trees, nearest neighbor or exemplar based learning, logistic regression, support vector machines
- Here we discuss support vector machines, logistic regression, and conditional random fields
- Then, we move to more advanced methods such as structured perceptrons, structured support vector machines and more general graphical models

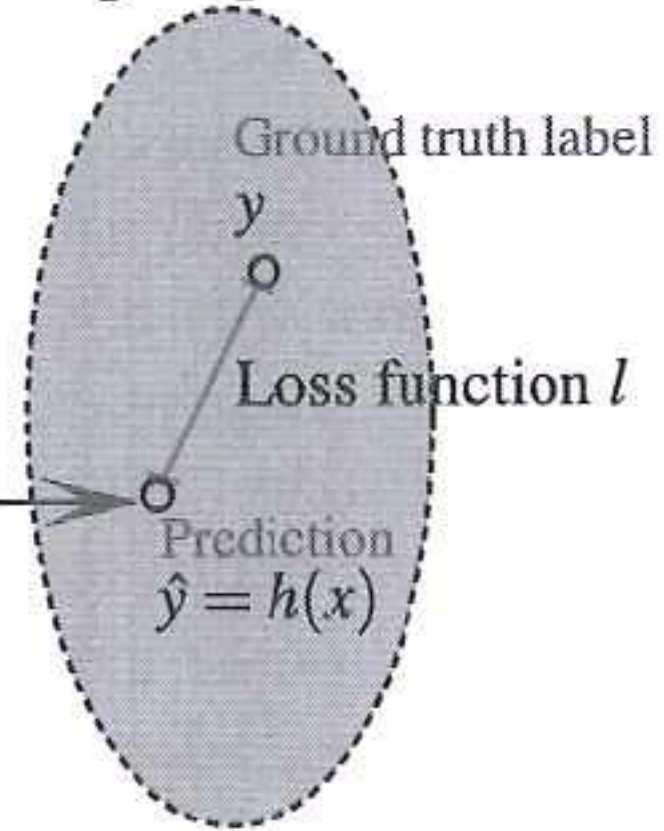
MACHINE LEARNING FRAMEWORK

- **Input space:** objects are represented as feature vectors
- **Output space:**
 - Regression: the space of real numbers
 - Classification: the set of discrete categories: $C = \{C_1, C_2, \dots, C_m\}$
- **Hypothesis space** = class of function mappings from the input space to the output space
- **To learn a good hypothesis:** in supervised learning a training set is used which contains a number of objects and their ground truth labels
- **Loss function:** to what degree the prediction generated by the hypothesis is in accordance with the ground truth label

Input Space X



Output Space Y



Hypothesis h

SUPPORT VECTOR MACHINE

- **Support vector machine:**
 - when two classes are linearly separable:
 - find a hyperplane in the p -dimensional feature space that best separates with **maximum margins** the positive and negative examples
 - maximum margins: with maximum Euclidean distance (= margin d) to the closest training examples (**support vectors**)
 - e.g., decision surface in two dimensions
 - idea can be generalized to examples that are not necessarily linearly separable and to examples that cannot be represented by linear decision surfaces

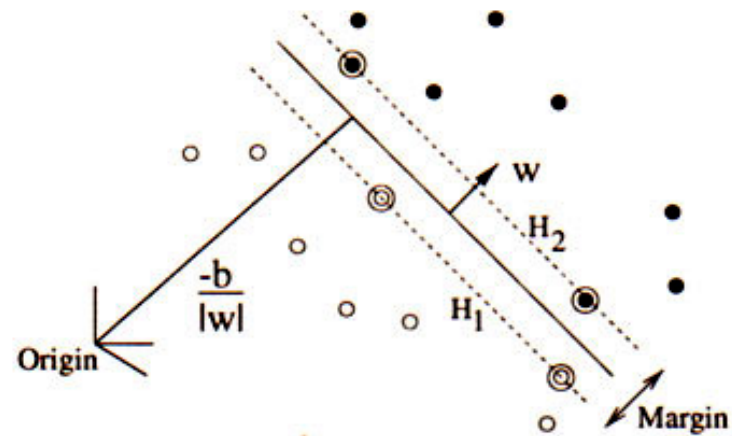


Figure 5. Linear separating hyperplanes for the separable case. The support vectors are circled.

[Burges 1998]

SUPPORT VECTOR MACHINE

- **Linear support vector machine:**

- case: **trained on data that are separable** (simple case)
- input is a set of n training examples:

$$S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$$

where $\mathbf{x}_i \in \mathfrak{R}^p$ and $y_i \in \{-1, +1\}$ indicating that \mathbf{x}_i is a negative or positive example respectively

- In case the data objects are not necessarily completely linearly separable (**soft margin SVM**):

$$\text{Minimize}_{\xi, \mathbf{w}, b} \langle \mathbf{w} \cdot \mathbf{w} \rangle + G \sum_{i=1}^n \xi_i^2$$

$$\text{Subject to } y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1 + \xi_i \geq 0, \quad i = 1, \dots, n$$

the amount of training error is measured using slack variables ξ_i the sum of which must not exceed some upper bound

where $\sum_{i=1}^n \xi_i^2$ = penalty for misclassification
 G = weighting factor

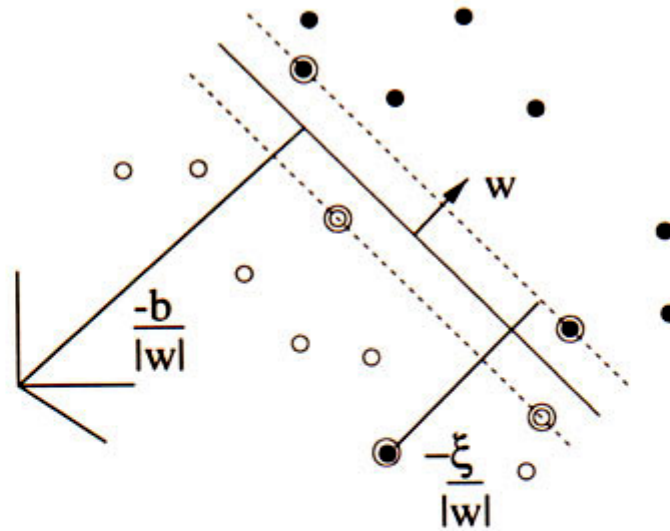


Figure 6. Linear separating hyperplanes for the non-separable case.

[Burges DMKD 1998]

A dual representation is obtained by introducing Lagrange multipliers λ_i , which turns out to be easier to solve:

$$\text{Maximize } W(\lambda) = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,j=1}^n \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \quad (1)$$

Subject to : $\lambda_i \geq 0$

$$\sum_{i=1}^n \lambda_i y_i = 0, \quad i = 1, \dots, n$$

Yielding the following decision function:

$$h(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$$

$$f(\mathbf{x}) = \sum_{i=1}^n \lambda_i y_i \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b \quad (2)$$

The decision function only depends on support vectors, i.e., for which $\lambda_i > 0$. Training examples that are not support vectors have no influence on the decision function

SUPPORT VECTOR MACHINE

- When classifying natural language data, it is not always possible to linearly separate the data: in this case we can map them into a feature space where they are linearly separable
- Working in a high dimensional feature space gives computational problems, as one has to work with very large vectors
- In the dual representation the data appear only inside inner products (both in the training algorithm shown by (1) and in the decision function of (2)): in both cases a kernel function can be used in the computations

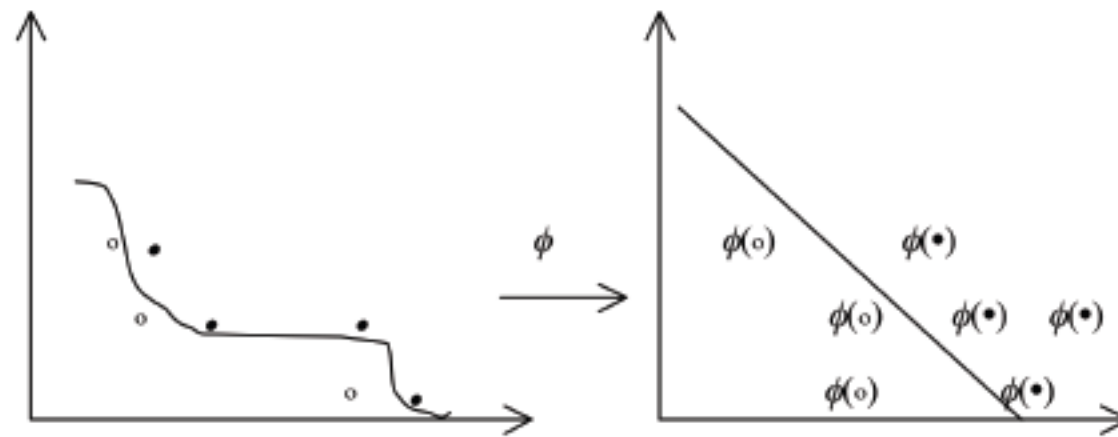


Fig. 5.2. A mapping of the features can make the classification task more easy (after Christianini and Shawe-Taylor 2000).

KERNEL FUNCTION

- A kernel function K is a mapping $K: S \times S \rightarrow [0, \infty]$ from the instance space of examples S to a similarity score:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$$

- In other words a kernel function is an inner product in some feature space
- The kernel function must be:
 - symmetric [$K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$]
 - positive semi-definite: if $\mathbf{x}_1, \dots, \mathbf{x}_n \in S$, then the $n \times n$ matrix G (*Gram matrix or kernel matrix*) defined by $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ is positive semi-definite*

* has non-negative eigenvalues

SUPPORT VECTOR MACHINE

- Typical kernel functions: linear (mostly used in text categorization), polynomial, radial basis function (RBF)
- We can define kernel functions that (efficiently) compare strings (**string kernel**) or trees (**tree kernel**)
- The decision function $f(\mathbf{x})$ we can just replace the dot products with kernels $K(\mathbf{x}_i, \mathbf{x}_j)$:

$$h(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$$

$$f(\mathbf{x}) = \sum_{i=1}^n \lambda_i y_i \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) \rangle + b$$

$$f(\mathbf{x}) = \sum_{i=1}^n \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

LINEAR REGRESSION

- Linear regression:

$$y = w_0 + \sum_{i=1}^N w_i \times f_i = w \cdot f$$

- Class membership:

$$p(y = \text{true} | \mathbf{x}) = \sum_{i=0}^N w_i \times f_i = w \cdot f \quad (3)$$

LINEAR REGRESSION

- Training of the model of (3):
 - By assigning each training example that belongs to the class the value $y = 1$, and the target value $y = 0$, if it is not
 - Train the weight vector to minimize the predictive error from 1 (for observations in the class) or 0 (for observations not in the class)
- Testing: dot product of the learned weight vector with the feature vector x of the new example
- But, result is not guaranteed to lie in $[0,1]$

LOGISTIC REGRESSION

- We predict a ratio of two probabilities as the log odds (or logit) function:

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right)$$

- **Logistic regression**: model of regression in which we use a linear function to estimate the logit of the probability

$$\ln\left(\frac{p(y = \text{true}|x)}{1-p(y = \text{true}|x)}\right) = w \cdot f$$

$$p(y = \text{true}|x) = \frac{e^{w \cdot f}}{1 + e^{w \cdot f}}$$

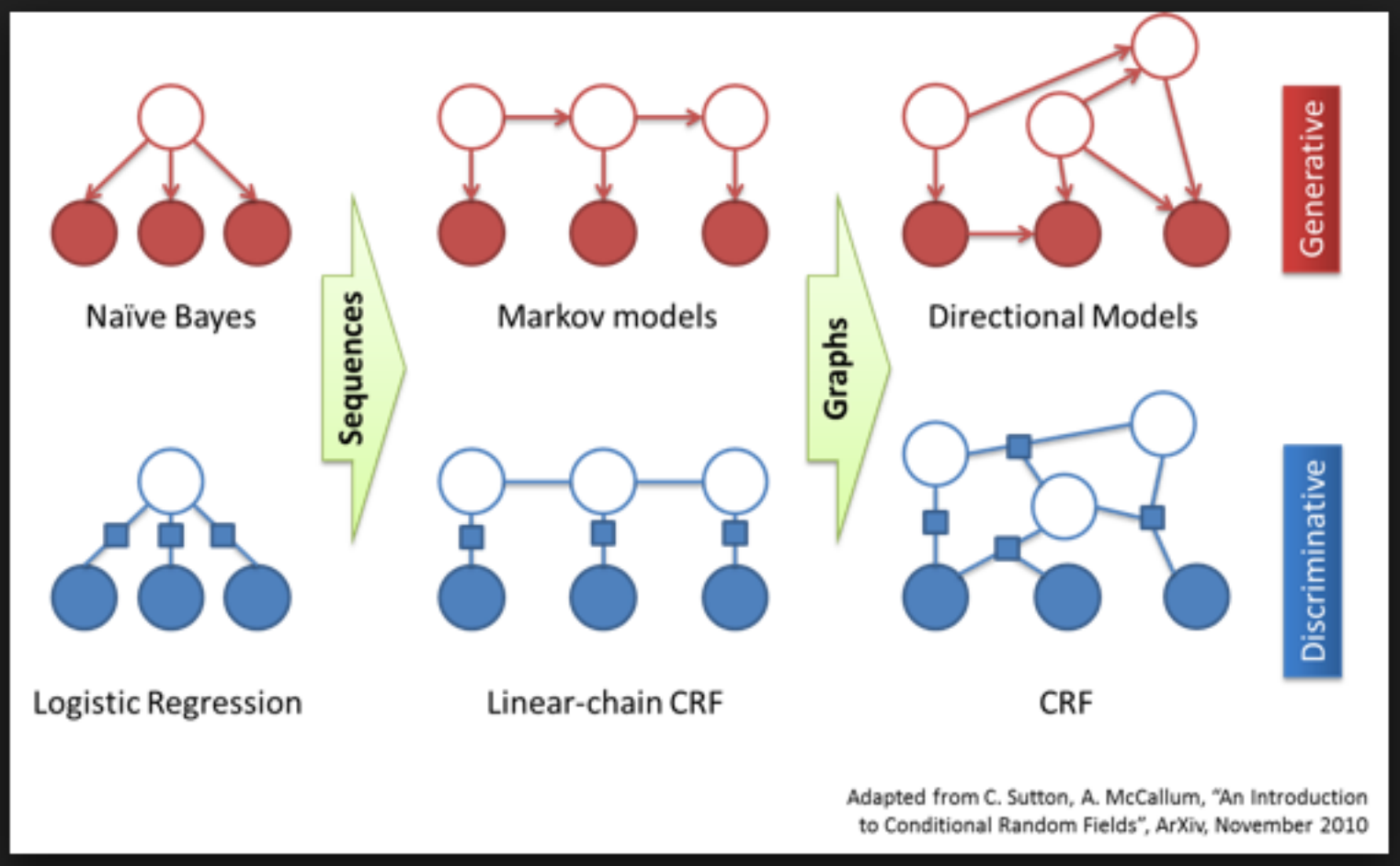
MULTINOMIAL LOGISTIC REGRESSION

- = **Maximum entropy classifier** (Maxent) deals with a larger number of classes: multinomial logistic regression
- Let there be C different classes: y_1, y_2, \dots, y_C
- We estimate the probability that y is a particular class y given N **feature functions** as:

$$p(y|\mathbf{x}) = \frac{1}{Z} \exp \sum_{i=0}^N w_i f_i$$
$$p(y|\mathbf{x}) = \frac{\exp \sum_{i=0}^N w_i f_i(y, \mathbf{x})}{\sum_{y' \in C} \exp \sum_{i=0}^N w_i f_i(y', \mathbf{x})}$$

- **Context dependent classification** = the class to which a feature vector is assigned depends on:
 - 1) the feature vector itself
 - 2) the values of other feature vectors and their class
 - 3) the existing relation among the various classes

- Examples:
 - conditional random field
 - structured output support vector machine



CONDITIONAL RANDOM FIELD

- **Linear chain conditional random field:**

- Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ be a random variable over data sequences to be labeled and Y a random variable over the corresponding label sequences
- All components y_i of Y are assumed to range over a finite label alphabet Σ
- We define $G = (V, E)$ to be an **undirected graph** such that there is a node $v \in V$ corresponding to each of the random variables representing an element y_v of Y
- If each y_v obeys the Markov property with respect to G , then the model (Y, X) is a conditional random field

CONDITIONAL RANDOM FIELD

- In an information extraction task, X might range over the words or constituents of a sentence/discourse, while Y ranges over the semantic/pragmatic classes to be recognized in these sentences/discourse
- **Template based or general CRF**: In theory the structure of graph G may be arbitrary: e.g., template based or general CRF, where you can define the dependencies in the Markov network or graph

[Lafferty et al. ICML 2001]

CONDITIONAL RANDOM FIELD

- To classify a new instance $P(Y | X)$ is computed as follows:

$$p(Y|X) = \frac{1}{Z} \exp\left(\sum_{j=1}^T \sum_{i=1}^k \lambda_i f_i(y_{j-1}, y_j, X, j)\right)$$

where

$f_i(y_{j-1}, y_j, X, j)$ = one of the k binary-valued feature functions

λ_i = parameter that models the observed statistics in the training examples

Z = normalizing constant

- The most probable label sequence Y^* for input sequence X is:

$$Y^* = \underset{Y}{\operatorname{argmax}} p(Y|X)$$

CONDITIONAL RANDOM FIELD

- **CRF training:**
 - Like for the Maxent model, we need numerical methods in order to derive λ_i
 - E.g., linear-chain CRF: variation of the Baum-Welch algorithm
 - In general CRFs we use approximate inference (e.g., Markov Chain Monte Carlo sampler)
- **Advantages and disadvantages:**
 - Very successful IE technique
 - Training is computationally expensive, especially when the graphical structure is complex

GLOBAL LEARNING

- Global or jointly recognizing several labels and their relationship
- Can be realized by:
 - Inferring a grammar (with rules) from data
 - Structured support vector machines
 - Graphical models (Markov random fields and Bayesian networks)

MODELS THAT JOINTLY LEARN

- The machine recognizes fragmentary pieces (e.g., names, facts) and the recognition of related fragments of text are often limited to the sentence level
- Emerging recognition of integrated understanding: e.g., in a discourse noun-phrase coreference resolution and entity recognition



**Human understanding of text:
inferencing,
connecting content**



[Wikipedia]

```

T
|--D
| |--: For these reasons, the Commission by a majority declares the application admissible,
|     without prejudging the merits.
|--A
|   |--A
|   | |--C
|   | |--: It follows that the application cannot be dismissed as manifestly ill-founded.
|   | |--A
|   |   |--P
|   |   |--: It considers that the applicant 's complaints raise serious issues of fact
|   |   |     and law under the convention, the determination of which should depend on
|   |   |     an examination of the merits.
|   |   |--P
|   |   |--: The Commission has taken cognizance of the submissions of the parties.
|   |--A
|   | |--C
|   | |--: In these circumstances, the Commission finds that the application cannot be
|   |     declared inadmissible for non-exhaustion of domestic remedies.
|   |--A
|   | |--P
|   | |--: The Commission recalls that article art. x of the convention only requires
|   |     the exhaustion of such remedies which relate to the breaches of the
|   |     convention alleged and at the same time can provide effective and sufficient
|   |     redress.
|   | |--P
|   | |--: The Commission notes that in the context of the section powers the
|   |     secretary of state has a very wide discretion.
|   | |--P
|   | |--: The Commission recalls that in the case of temple v. the united kingdom
|   |     no. x dec. d.r. p.
|   | |--P
|   | |--: The Commission held that recourse to a purely discretionary power on
|   |     the part of the secretary of state did not constitute an effective
|   |     domestic remedy.
|   | |--: The Commission finds that the suggested application for discretionary
|   |     relief in the instant case cannot do so either.

```

Fig. 6: Output of the automatic system: small fragment of the argumentation tree-structure of a document

INFERRING A GRAMMAR WITH RULES FROM DATA

- Can be done manually (cf. PhD of Raquel Mochales Palau)
- Can be learned from annotated data
- Could be learned from a very large unannotated corpus, but very difficult if grammar is complex

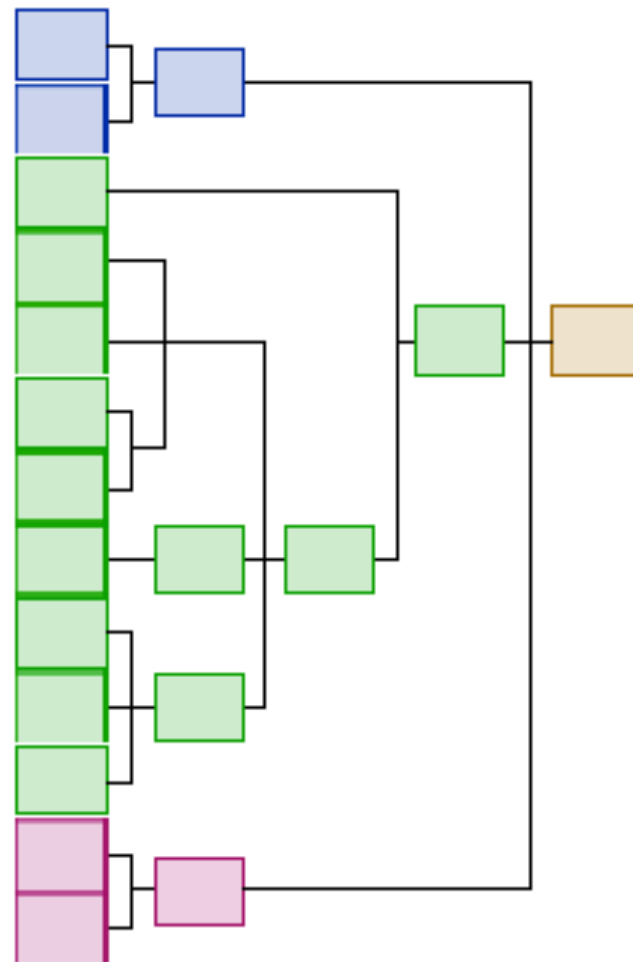


Figure 1.1: Reasoning structure of the legal case in Appendix A. Each block is a sentence of the legal case. There are 3 arguments (blue, green and red) that justify the final decision (brown). The contents of each argument and the final decision can be seen in detail in Figures 1.2, 1.3, 1.4 and 1.5

[PhD thesis Raquel Mochales Palau 2011]

Experiments with decisions of the European Court of Human Rights (ECHR)

$$\begin{aligned}
 T &\Rightarrow A^+ D \\
 A &\Rightarrow \{A^+ C | A^+ C n P^+ | C n s | A^+ s r_c C | P^+\} \\
 D &\Rightarrow r_c f \{v_c s | \cdot\}^+ \\
 P &\Rightarrow \{P_{verb} P | P_{art} | P P_{sup} | P P_{ag} | s P_{sup} | s P_{ag}\} \\
 P_{verb} P &= s v_p s \\
 P_{art} &= s r_{art} s \\
 P_{sup} &= \{r_s\} \{s | P_{verb} P | P_{art} | P_{sup} | P_{ag}\} \\
 P_{ag} &= \{r_a\} \{s | P_{verb} P | P_{art} | P_{sup} | P_{ag}\} \\
 C &= \{r_c | r_s\} \{s | P_{verb} P | C\} \\
 C &= s^+ v_c s
 \end{aligned}$$

Fig. 5: Context-free grammar used for argumentation structure detection and proposition classification

[Mochales & Moens AI & Law 2011]

Table 9: Terminal and non-terminal symbols from the context-free grammar used in the argumentation structure detection

T	General argumentative structure of legal case.
A	Argumentative structure that leads to a final decision of the factfinder $A = \{a_1, \dots, a_n\}$, each a_i is an argument from the argumentative structure.
D	The final decision of the factfinder $D = \{d_1, \dots, d_n\}$, each d_i is a sentence of the final decision.
P	One or more premises $P = \{p_1, \dots, p_n\}$, each p_i is a sentence classified as premise.
P_{CG}	Premise with at least one contrast rhetorical marker.
P_{art}	Premise with at least one article rhetorical marker.
P_{sup}	Premise with at least one support rhetorical marker.
$P_{\text{verb}P}$	Premise with at least one verb related to a premise.
C	Sentence with a conclusive meaning.
n	Sentence, clause or word that indicates one or more premises will follow.
s	Sentence, clause or word neither classified as a conclusion nor as a premise ($s \in \{C P\}$).
r_c	Conclusive rhetorical marker (e.g. therefore, thus, ...).
r_s	Support rhetorical marker (e.g. moreover, furthermore, also, ...).
r_a	Contrast rhetorical marker (e.g. however, although, ...).
r_{art}	Article reference (e.g. terms of article, art. para. ...).
v_p	Verb related to a premise (e.g. note, recall, state, ...).
v_c	Verb related to a conclusion (e.g. reject, dismiss, declare, ...).
f	The entity providing the argumentation (e.g. court, jury, commission, ...).

INFERRING A GRAMMAR WITH RULES FROM DATA

- Works well (see further the results)
- A deterministic grammar might overfit the data it is constructed from
- A probabilistic grammar needs annotated data
- If we have annotated data we can learn the grammar

INTERMEZZO: SPATIAL RELATION

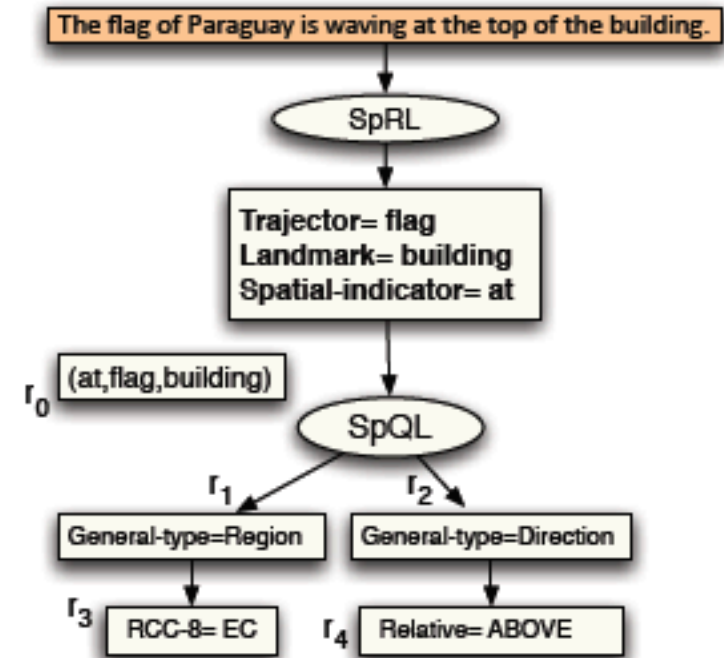
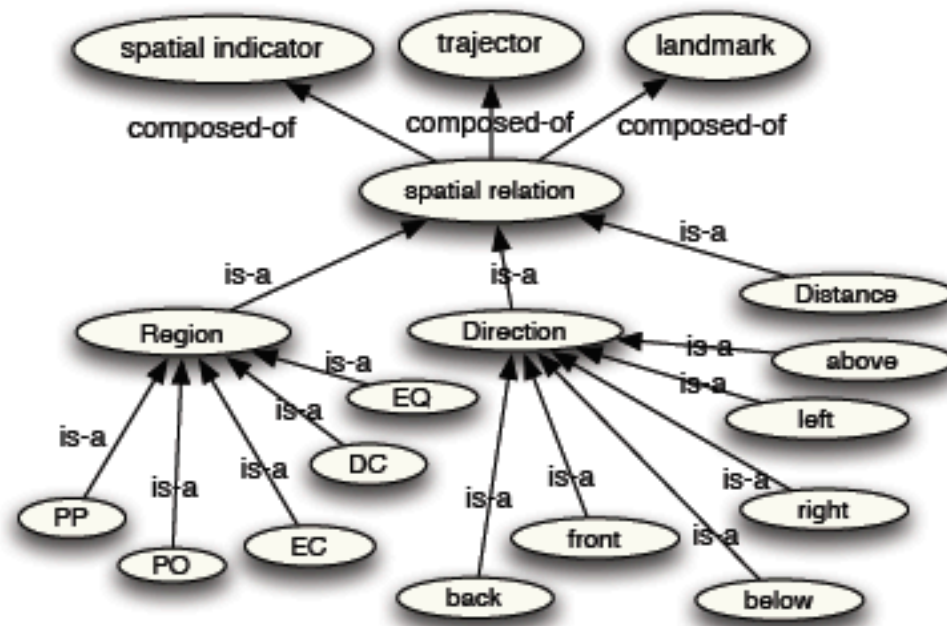


Figure 1. (a) The spatial ontology.

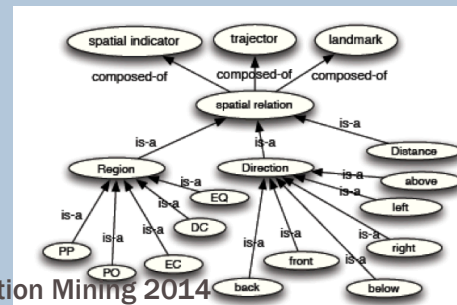
(b) Example sentence and the recognized spatial concepts.

The goal is to jointly assign the labels of the ontology to a text item

JOINT MACHINE LEARNING

- **Joint or global learning** \neq local learning of independent classifiers
 - Independent classifiers and combination of results (e.g., based on integer linear programming)
 - Joint training:
 - 1 classification model for the global structure: cf. CRF
 - Output is = structure (e.g., spatial ontology)

[PhD of Parisa Kordjamshidi 2013]
[Kordjamshidi & Moens Journal of Web Semantics 2014]



OUTPUT

- Output variables = labels in the structure

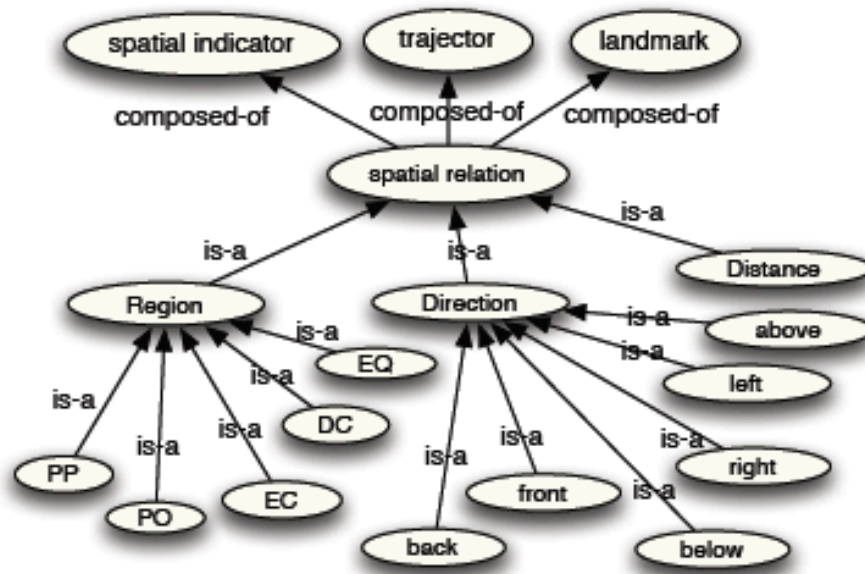
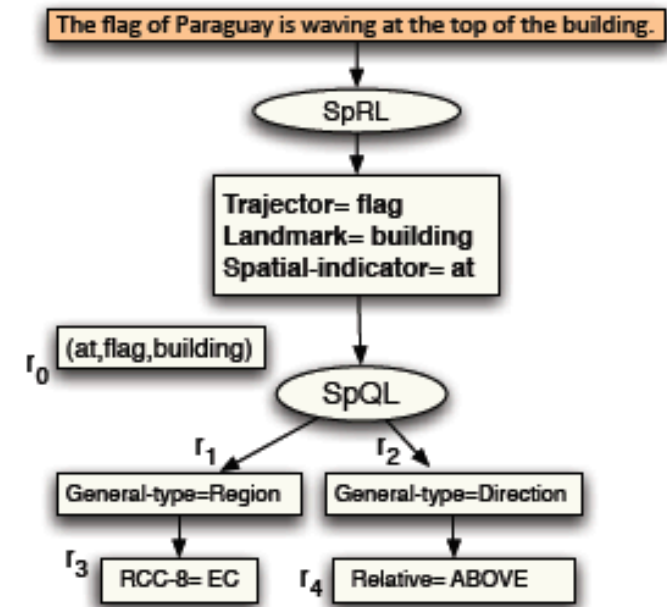


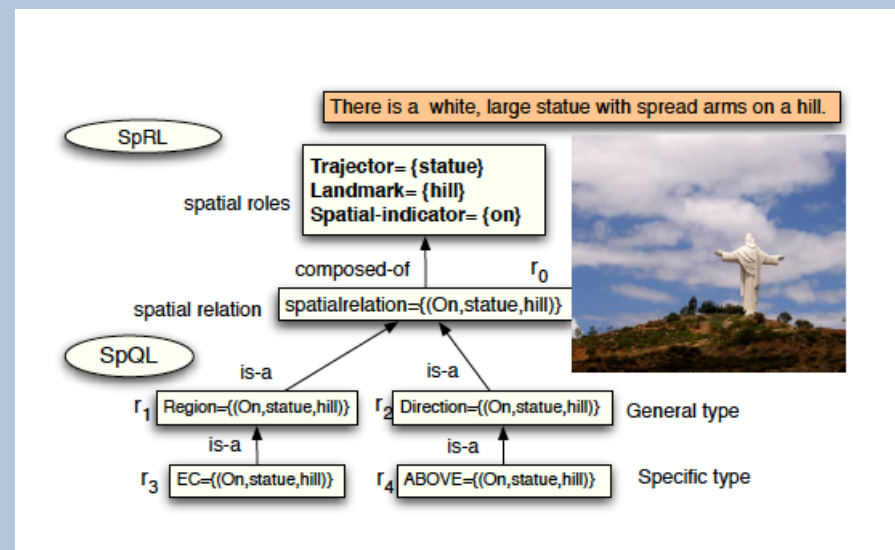
Figure 1. (a) The spatial ontology.



(b) Example sentence and the recognized spatial concepts.

INPUT

- Object to which the classification model is applied: e.g., sentence (in our case), paragraph, full document, ...
- Is usually composed of different **input components**: single words, phrases, ... depending on the type of text snippet to which a label will be assigned



FEATURE FUNCTIONS

- Each input component is assigned a set of features: e.g., lexical, syntactic, discourse distance, ...
- Feature functions link an input component with a possible label (notion of feature templates)
- Each feature function will receive a weight during training
- A feature template groups a set of feature functions => block of corresponding weights W_i

OBJECTIVE FUNCTION

- The main objective discriminant function

$$g(x, y; W) = \langle W, f(x, y) \rangle$$

is a linear function in terms of the combined feature representation associated with each candidate input component and an output label according to the template (Ψ) specifications

- Can be written in terms of the instantiations of the templates and their related blocks of weights W_p

TRAINING OF THE MODEL

- A popular discriminative training approach is to minimize the following convex upper bound of the loss function over the N training data:

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \forall i, \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_i : \\ & \mathbf{w}^T \Psi(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^T \Psi(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \end{aligned}$$

- Training with the most violated constraints/outputs (\mathbf{y}) per training example
- In the experiments: structured support vector machines (SSVM), structured perceptrons and averaged structured perceptrons

CONSTRAINTS

$$sp_i + nsp_i = 1$$

$$sp_i tr_j + sp_i lm_j + sp_i nrol_j = 1$$

$$sp_i tr_j - sp_i \leq 0, \quad sp_i lm_j - sp_i \leq 0$$

$$sp_i - \sum_j (sp_i tr_j) \leq 0, \quad sp_i - \sum_j (sp_i lm_j) \leq 0$$

$$sp_i tr_j + sp_i lm_j \leq 1$$

$$\sum_i (sp_i tr_j) \leq 1, \quad \sum_i (sp_i lm_j) \leq 1$$

$$sp_i tr_j lm_k r_\gamma - sp_i tr_j lm_k r_{\gamma'} \leq 0, \quad \forall \gamma < \gamma' \quad \gamma, \gamma' \in \mathcal{H}$$

$$\sum_{\gamma \in \mathcal{H}_{leafs}} r_\gamma(x_i, x_j, x_k) \geq r_0(x_i, x_j, x_k)$$

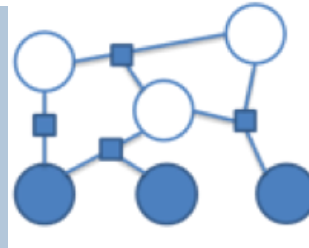
$$\sum_{\gamma \in QSR_h} sp_i tr_j lm_k r_\gamma \leq 1, \quad \forall h, \quad \forall QSR_h \subset \mathcal{H}_{leafs}$$

Constraints are linear and variables take the form of integers

Constraints are applied:
during training:
finding the most violated outputs
and/or
during testing

GRAPHICAL MODELS IN GENERAL

- E.g., Markov random fields



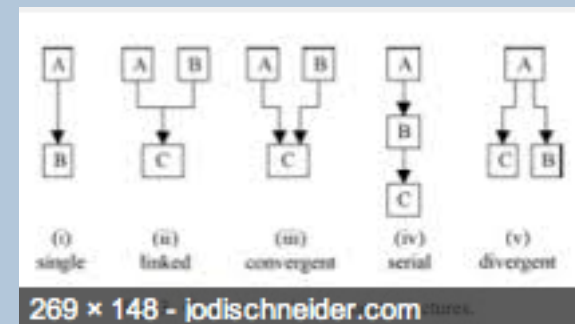
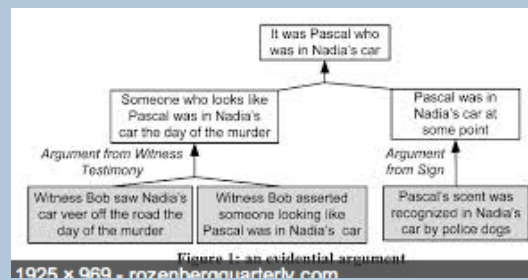
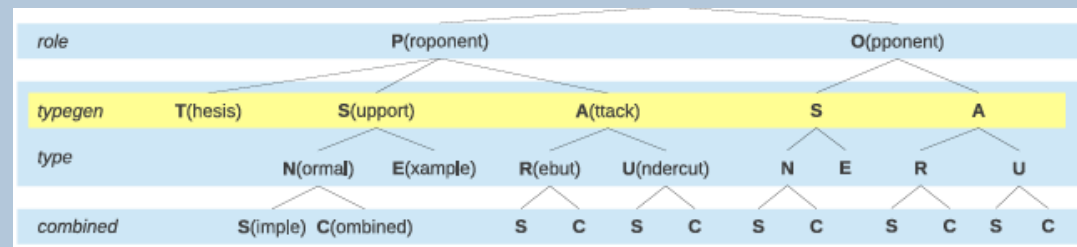
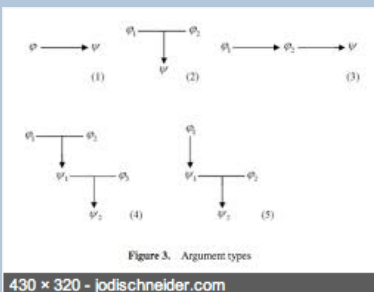
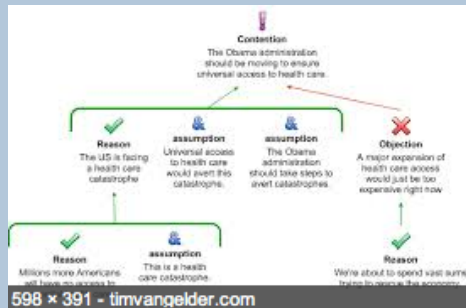
- Allow using rules as features for which the weight is trained on the annotated data
- Concern: the computational complexity

JOINT RECOGNITION OF A CLAIM AND ITS COMPOSING ARGUMENTS

- Structured learning: modeling of interdependence among output labels:
 - Generalized linear models, e.g., structured support vector machines and structured perceptrons [Tsochantaridis et al. JMLR 2006]
 - Probabilistic graphical models [Koller and Friedman 2009]
- The interdependencies between output labels and other background knowledge can be imposed using constraint optimization techniques during prediction and training
 - Cf. recent work on structure analysis of scientific documents [Guo et al. NAACL-HLT 2013]

OTHER ARGUMENTATION STRUCTURES

- Or to the Toulmin model or the many different argumentation schemes/**structures** discussed in Douglas Walton (1996). *Argumentation Schemes for Presumptive Reasoning*. Mahwah, New Jersey: Lawrence Erlbaum Associates
- Work of Prakken, Gordon, Bench-Capon, Atkinson, Wyner, Schneider, ...

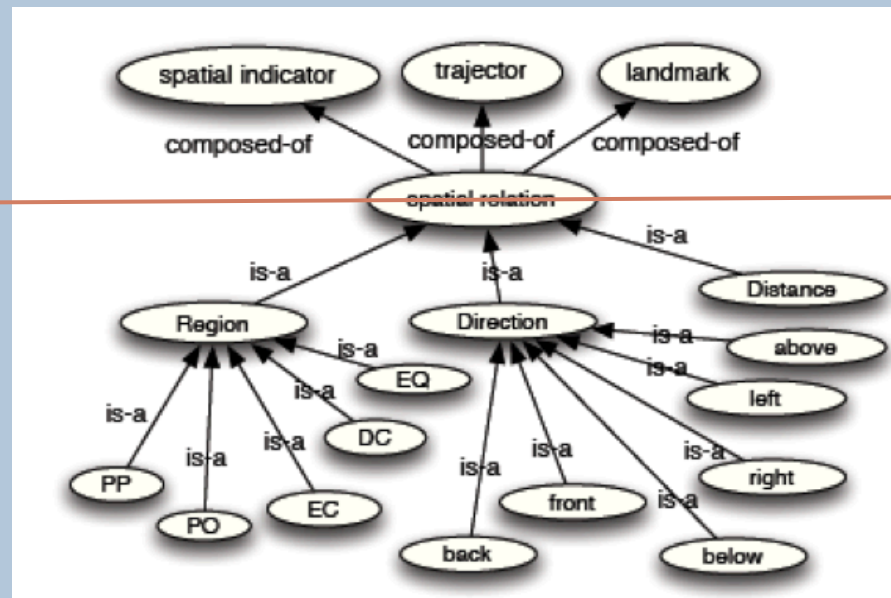


DECOMPOSITIONS

- **Complex graphical structures:** considering the interdependencies and structural constraints over the output space easily leads to intractable training and prediction situations:
 - Models for decomposition, communicative inference, message passing, ...
 - A current research topic in machine learning

DECOMPOSITIONS

- Breaking the structured model in two or more pieces:
 - Build a model for each piece
 - Possibly: Iteratively improve each model by communicating between the pieces



FEATURES REVISITED

- Argumentation mining

On the other hand the court notes that there are substantial delays attributable to the authorities

In particular in the first set of proceedings there is a period of inactivity of more than two years ...

In the second set of proceedings there is a period of inactivity of some three years

Premises

The court cannot find that the government has given sufficient explanation for these delays that occurred

Conclusion

FEATURES REVISITED

- Because we input candidate arguments and their candidate components:
 - We can describe the component with different features than the ones used for describing the full argument
 - E.g., textual entailment relationships can be used to describe the full argument

- Our argumentation mining machine only uses information resided in the texts



[Wikipedia]

- **Human understanding of text: humans connect to their world/ domain knowledge**

■ The discourse structure is often signaled by typical keywords (e.g., in conclusion, however, ...), but often this is not the case

■ Humans who understand the meaning of the text can infer whether a claim is a plausible conclusion given a set of premises, or a claim rebuts another claim

⇒ Background or domain knowledge that makes a certain discourse relation valid

⇒ Background or domain knowledge that an argumentation mining tool should also acquire: how?

■ Work on **textual entailment** : [Cabrio & Villata 2012], event causality: [Xuan Do et al. EMNLP 2011], ...

TEXTUAL ENTAILMENT

- Textual entailment: recognize, given two text fragments whether one text can be inferred (entailed) from the other
- Has been studied widely in computational linguistics and the machine learning communities (e.g., Pascal recognizing textual entailment challenge)

Example 1.

T1: Research shows that drivers speaking on a mobile phone have much slower reactions in braking tests than non-users, and are worse even than if they have been drinking.

H: The use of cell-phones while driving is a public hazard.

Example 2 (Continued).

T2: Regulation could negate the safety benefits of having a phone in the car. When you're stuck in traffic, calling to say you'll be late can reduce stress and make you less inclined to drive aggressively to make up lost time.

H: The use of cell-phones while driving is a public hazard.

TEXTUAL ENTAILMENT

- Most of the work in textual entailment: approaches of distance computation between the texts (e.g. edit distances, similarity metrics, kernels):
 - E.g., EDITS system (Edit Distance Textual Entailment Suite), an open-source software package for textual entailment: <http://edits.fbk.eu/>

ENTAILMENT IN ARGUMENTATION

Example 3 (Continued).

T3: If one is late, there is little difference in apologizing while in their car over a cell phone and apologizing in front of their boss at the office. So, they should have the restraint to drive at the speed limit, arriving late, and being willing to apologize then; an apologetic cell phone call in a car to a boss shouldn't be the cause of one being able to then relax, slow-down, and drive the speed-limit.

T2 → HI: Regulation could negate the safety benefits of having a phone in the car. When you're stuck in [...]

[Cabrio & Vilata ACL 2012]

TE provides techniques to detect both the argument components, and the kind of relation underlying them:
Or an entailment or a contradiction is detected

ENTAILMENT IN ARGUMENTATION

- Similarity measures are rough approaches
- Very difficult to acquire automatically the background knowledge needed for the entailment:
- => process that takes years for legal professionals

■ Part 3: Some applications

ARGUMENTATION MINING OF LEGAL CASES

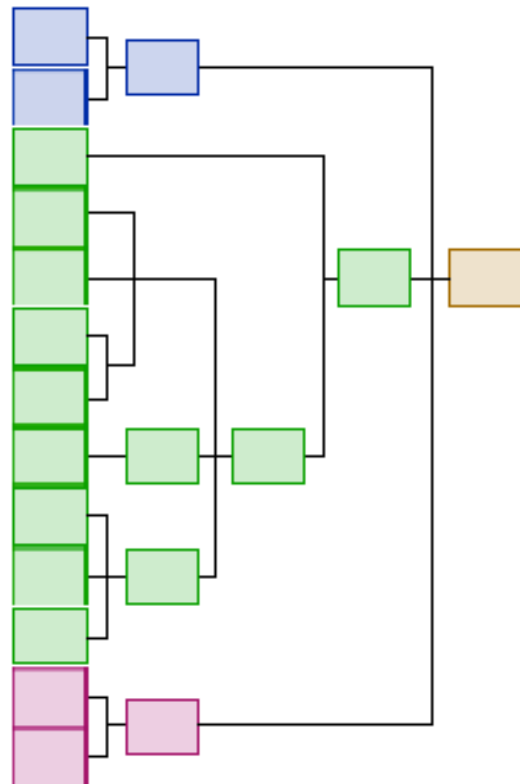


Figure 1.1: Reasoning structure of the legal case in Appendix A. Each block is a sentence of the legal case. There are 3 arguments (blue, green and red) that justify the final decision (brown). The contents of each argument and the final decision can be seen in detail in Figures 1.2, 1.3, 1.4 and 1.5

Cases of the European Court of Human Rights

[PhD thesis Raquel Mochales Palau 2011]

Features of classifier:

Clauses described by unigrams, bigrams, adverbs, legal keywords, word couples over adjacent clauses, ...

Table 7: Results from the classification of *Conclusions* in the ECHR

Classifier Combination	Precision	Recall	F-Measure
Max.Ent. and Support Vector Machine	77.49	60.88	74.07
Context-free Grammar	61.00	75.00	67.27

Table 8: Results from the classification of *Premises* in the ECHR

Classifier Combination	Precision	Recall	F-Measure
Maxt.Ent. and Support Vector Machine	70.19	66.16	68.12
Context-free Grammar	59.00	71.00	64.03

Context free grammar allows also to recognize the full argumentation structure: accuracy: 60%

[Mochales & Moens AI & Law 2011]

SUPPORT FOR ONLINE USER COMMENTS

- Online user comments contain arguments with appropriate or missing justification
- [Park & Cardie FWAM 2014] classify comments into classes such as UNVERIFIABLE, VERIFIABLE NON-EXPERIENTIAL, VERIFIABLE EXPERIENTIAL

	#	proposition
VERIFEX P	1	I've been a physician for 20 years.
	2	My son has hypolycemia.
	3	They flew me to NY in February.
	4	The flight attendant yelled at the passengers.
VERIFNON	5	<i>They can have inhalation reactions.</i>
	6	<i>since they serve them to the whole plane.</i>
	7	<i>Peanuts do not kill people.</i>
	8	<i>Clearly, peanuts do not kill people.</i>
	9	<i>I believe peanuts do not kill people.</i>
	10	<i>The governor said that he enjoyed it.</i>
	11	<i>food allergies are rare</i>
	12	<i>food allergies are seen in less than 20% of the population</i>
UNVERIF	13	Again, keep it simple.
	14	Banning peanuts will reduce deaths.
	15	I enjoy having peanuts on the plane.
	16	others are of uncertain significance
	17	banning peanuts is a slippery slope
NONARG	18	Who is in charge of this?
	19	I have two comments
	20	http://www.someurl.com
	21	Thanks for allowing me to comment.
	22	- Mike

Table 1: Example Sentences.

SUPPORT FOR ONLINE USER COMMENTS

- Features: n-grams, POS tags, present in core or accessory clause, sentiment clue, speech event anchors, imperative expression count, emotion expression count, tense count, person count

	VERIF _{NON}	VERIF _{EXP}	UNVERIF	Total
Train	987	900	4459	6346
Test	370	367	1687	2424
Total	1357	1267	6146	8770

Table 4: # of propositions in Train and Test Set

Feature Set	UNVERIF vs All			VERIF _{NON} vs All			VERIF _{EXP} vs All			Average F ₁	
	Pre.	Rec.	F ₁	Pre.	Rec.	F ₁	Pre.	Rec.	F ₁	Macro	Micro
<i>UNI(base)</i>	85.24	79.43	82.23	42.57	51.89	46.77	61.10	66.76	63.80	64.27	73.31
<i>UNI+BI</i>	82.14	89.69*	85.75*	51.67*	37.57	43.51	73.48*	62.67	67.65*	65.63	77.64*
<i>VER</i>	88.52*	52.10	65.60	28.41	61.35*	38.84	42.41	73.02*	53.65	52.70	56.68
<i>EXP</i>	82.42	4.45	8.44	20.92	76.49*	32.85	31.02	82.83*	45.14	28.81	27.31
<i>VER+EXP</i>	89.40*	49.50	63.72	29.25	71.62*	41.54	50.00	79.56*	61.41	55.55	57.43
<i>UNI+BI+VER+EXP</i>	86.86*	83.05*	84.91*	49.88*	55.14	52.37*	66.67*	73.02*	69.70*	68.99*	77.27*

Table 3: Three class classification results in % (Crammer & Singer's Multiclass SVMs)

RECOGNIZING ARGUMENTS IN ONLINE DISCUSSIONS

- Boltužic & Šnajder FWAM 2014 identify properties of comment-argument pairs

Label	Description: Comment...
A	...explicitly attacks the argument
a	...vaguely/implicitly attacks the argument
N	...makes no use of the argument
s	...vaguely/implicitly supports the argument
S	...explicitly supports the argument

Table 2: Labels for comment-argument pairs in the COMARG corpus

Topic	Labels					Total
	A	a	N	s	S	
UGIP	48	86	691	58	130	1,013
GM	89	73	849	98	176	1,285
UGIP+GM	137	159	1,540	156	306	2,298

Table 5: Distribution of labels in the COMARG corpus

RECOGNIZING ARGUMENTS IN ONLINE DISCUSSIONS

- Features: entailment features (TE): from pretrained entailment decision algorithms (which a.o. use WordNet, VerbOcean); semantic text similarity features (STS) and stance alignment feature (SA) with stance known a priori
- Multiclass classification with support vector machine

Model	A-a-N-s-S		Aa-N-sS		A-N-S	
	UGIP	GM	UGIP	GM	UGIP	GM
MCC baseline	68.2	69.4	68.2	69.4	79.5	76.6
BoWO baseline	68.2	69.4	67.8	69.5	79.6	76.9
TE	69.1	81.1	69.6	72.3	80.1	73.4
STS	67.8	68.7	67.3	69.9	79.2	75.8
SA	68.2	69.4	68.2	69.4	79.5	76.6
STS+SA	68.2	69.5	67.5	68.7	79.6	76.1
TE+SA	68.9	72.4	71.0	73.7	81.8	80.3
TE+STS+SA	70.5	72.5	68.9	73.4	81.4	79.7

Table 7: Argument recognition F1-score (separate models for UGIP and GM topics)

Model	UGIP → GM		GM → UGIP	
	A-a-N-s-S	Aa-N-sS	A-a-N-s-S	Aa-N-sS
STS+SA	69.4	69.4	68.2	68.2
TE+SA	72.6	73.5	70.2	71.2
STS+TE+SA	71.5	72.2	68.2	69.6

Table 8: Argument recognition F1-score on UGIP and GM topics (cross-topic setting)

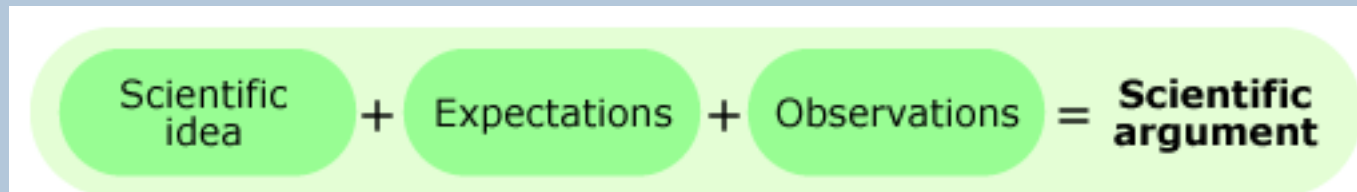
Boltužic & Šnajder FWAM 2014

ARGUMENT ENRICHED OPINION MINING

- Opinion mining: finding arguments and counter arguments for an opinion expressed:
 - Find support for the opinion, explain the opinion
 - *An opinion, whether it is grounded in fact or completely unsupportable, is an idea that an individual or group holds to be true. An opinion does not necessarily have to be supportable or based on anything but one's own personal feelings, or what one has been taught. An argument is an assertion or claim that is supported with concrete, real-world evidence.*
[<http://wiki.answers.com>]

ARGUMENT MINING IN THE SCIENTIFIC LITERATURE

- Mining of the supporting evidence of claims in scientific publications and patents and their visualization for easy access




[http://undsci.berkeley.edu/article/howscienceworks_07]

ARGUMENT MINING IN THE DIGITAL HUMANITIES

- Digital humanities: finding and comparing the arguments that politicians use in their speeches:
 - *Then that little man in black there, he says women can't have as much rights as men, 'cause Christ wasn't a woman! Where did your Christ come from? Where did your Christ come from? From God and a woman! Man had nothing to do with Him. [Sojourner Truth (1797-1883): Ain't I A Woman?, Delivered 1851, Women's Convention, Akron, Ohio]*

ANNOTATED DATA

- **The Araucaria corpus** (constructed by Chris Reed at the University of Dundee, 2003) now extended to **AIF-DB**
- The **ECHR corpus** annotated by legal experts in 2006 under supervision of Raquel Mochales Palau:
 - 25 legal cases
 - 29 admissibility reports
 - 12.904 sentences, 10.133 non-argumentative and 2.771 argumentative, 2.355 premises and 416 conclusions
- Plans to build corpus of **biomedical genetics research** literature [Green FWAM 2014]
- Several smaller corpora described in FWAM 2014
- ...



- **Part 4: Conclusions and thoughts for future research**

CONCLUSIONS

- Argumentation mining: novel and promising research domain
- Potential of joint learning of an argumentation structure integrating known interdependencies between the structural components in the argumentation and expert knowledge
- Potential of better textual entailment techniques
- Numerous interesting applications of the technology !



THOUGHTS FOR FUTURE RESEARCH

■ ?

■ **ISCH COST Action IS1312**

Structuring Discourse in Multilingual Europe (TextLink)

http://www.cost.eu/domains_actions/isch/Actions/IS1312

<http://textlinkcost.wix.com/textlink>