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

[Diagram of argumentative structure]

[Image of business people discussing]

Tutorial Argumentation Mining 2014
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

ARGUMENTATION

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

- Aristotle's (4th century BC) logical works: *Organon*
- George Pierce Baker, *The Principles of Argumentation*, 1895
- Chaïm Perelman describes techniques of argumentation used by people to obtain the approval of others for their opinions: *Traité de l'argumentation – la nouvelle rhétorique*, 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.

We find argumentation in:

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

TODAY
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
- ...

Tutorial Argumentation Mining 2014
ARGUMENTATIVE ZONING

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

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...)
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)

Distributional Clustering of English Words

Fernando Penedo
Youmeh Tshibi
Lillian Lee

Abstract

We introduce and experimentally evaluate a method for automatically clustering meaningful words in a text dividing the text into discourse segments or zones that each play a specific rhetoric role in a text. Our method is based on the observation that words with similar contexts also occur in the same discourse segments or zones. We evaluate our method on 837 abstracts and find that it improves the accuracy of automatic clustering by up to 50%.

Problem Setting

In this work, we will create two new word clusters, CRITICAL and COMMON, to better capture the different roles played by words in a text. We will also introduce a new measure, the weight of a word, which is a measure of the importance of a word in a text. We will use these new clusters and measure to improve the accuracy of automatic clustering in this domain.

[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

[Teufel, S. & Moens, M. ACL 1999]
[Teufel, S. & Moens, M. EMNLP 2000]
[Hachey, B. & Grover, C. ICAIL 2005]
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]
The Court notes that this complaint is not manifestly ill-founded within the meaning of Article 35 § 3 of the Convention.

It further notes that it is not inadmissible on any other grounds.

It must therefore be regarded as admissible.

Figure 1.2: Closer view 1st Argument

Article 41 of the Convention provides: "If the Court finds that there has been a violation of the Convention or the Protocols thereto, and if the internal law of the High Contracting Party concerned allows only partial reparation to be made, the Court shall, if necessary, afford just satisfaction to the injured party."

The applicant has not filed a claim for just satisfaction.

Accordingly, the Court considers that no award can be made under this provision.

Figure 1.3: Closer view 2nd Argument

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

[PhD thesis Raquel Mochales Palau 2011]
Figure 1.5: Closer view 3rd Argument

... as provided in Article 6 § 1 of the Convention, which reads as follows: In the determination of his civil rights and obligations,... everyone is entitled to fair... hearing within reasonable time... by(a)... tribunal.

The Court reiterates that the reasonableness of the length of proceedings must be assessed in the light of,... particularly the complexity of the case, the conduct of the applicant and of the relevant authorities.

The Court considers that the present proceedings... were not particularly complex.

As regards the conduct of the applicant... applicants cannot be blamed for making full use of the remedies available to them.

However, an applicant’s behaviour... whether or not the reasonable time referred to in Article 6 § 1... In the present case, the Court acknowledges that the applicant had filed numerous requests...

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...

Although such conduct contributed..., it is not sufficient to explain the length of the extensive proceedings.

The Court therefore finds that the overall length of the proceedings cannot be regarded as "reasonable".

Accordingly, there has been a violation of Article 6 § 1 of the Convention.
[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
For these reasons, the Commission by a majority declares the application admissible, without prejudging the merits.

It follows that the application cannot be dismissed as manifestly ill-founded.

The Commission has taken cognizance of the submissions of the parties.

In these circumstances, the Commission finds that the application cannot be declared inadmissible for non-exhaustion of domestic remedies.

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.

The Commission notes that in the context of the section powers the secretary of state has a very wide discretion.

The Commission recalls that in the case of temple v. the united kingdom no. x dec. d.r. 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

[Mocharles & Moens AI & Law 2011]

Tutorial Argumentation Mining 2014
Part 2: Introducing current methods
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
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 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
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
Examples $x_i$ classified with labels $y_i = \text{training set}$

Classification function or rules are learned

New instances to be classified $= \text{test set}$
In classification: given inputs $x$ and their labels $y$:

- **Generative classifier** learns a model of the joint probability $p(x,y) = p(y) p(x|y)$ and then condition on the observed features $x$, thereby deriving the class posterior $p(y|x)$ and selects the most probable $y$ for $x$

- Generative classifier: since it specifies how to generate the observed features $x$ for each class $y$

- E.g.,
  - Naive Bayes, hidden Markov model
\textbf{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)
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
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 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)
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
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**:
    \[
    x = [x_1, x_2, ..., x_p]^T
    \]
    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
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…
FEATURE VECTORS FOR AN EXAMPLE TEXT

A Java Applet that scans Java Applets

- Binary values, based on lower-cased words:
  
  \[
  \text{[a: 1, apple: 0, applet: 1, applets: 1, ..., java: 1, ...]}
  \]

- Remove stopwords:
  
  \[
  \text{[apple: 0, applet: 1, applets: 1, ..., java: 1 ...]}
  \]

- Numeric value: based on text term frequency (tf):
  
  \[
  \text{[apple: 0, applet: 1, applets: 1, ..., java: 2 ...]}
  \]

- Numeric value: based on text term frequency of lower cased n-grams (tf):
  
  \[
  \text{[aa: 0, a_a: 2, a_b: 0, ...]}
  \]

- Numeric attribute value based on latent semantic indexing:
  
  \[
  \text{[F1: 0.38228938, F2: 0.000388, F3: 0.201033, ...]}
  \]

- ...

Tutorial Argumentation Mining 2014

38
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
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: \( C = \{C_1, C_2, \ldots, C_m\} \)
- **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
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]
**Linear support vector machine:**

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

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

where $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$ indicating that $x_i$ is a negative or positive example respectively
In case the data objects are not necessarily completely linearly separable (soft margin SVM):

Minimize \( \xi, w, b \) \( \langle w \cdot w \rangle + G \sum_{i=1}^{n} \xi_i^2 \)

Subject to \( y_i(\langle w \cdot x_i \rangle + b) - 1 + \xi_i \geq 0 \), \( i = 1, \ldots, n \)

the amount of training error is measured using slack variables \( \xi_i \) the sum of which must not exceed some upper bound

where \( \sum_{i=1}^{n} \xi_i^2 = \text{penalty for misclassification} \)

\( G = \text{weighting factor} \)
Figure 6. Linear separating hyperplanes for the non-separable case.

[Burges DMKD 1998]
A dual representation is obtained by introducing Lagrange multipliers $\lambda_i$, which turns out to be easier to solve:

Maximize $W(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \langle x_i \cdot x_j \rangle$ \hspace{1cm} (1)

Subject to: $\lambda_i \geq 0$

$$\sum_{i=1}^{n} \lambda_i y_i = 0 \hspace{0.5cm}, \hspace{0.5cm} i = 1,\ldots,n$$
Yielding the following decision function:

\[ h(x) = \text{sign}(f(x)) \]

\[ f(x) = \sum_{i=1}^{n} \lambda_i y_i \langle x_i \cdot x \rangle + b \]  

(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.
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).
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:

$$K(x_i, x_j) = \langle \phi(x_i) \cdot \phi(x_j) \rangle$$

In other words a kernel function is an inner product in some feature space.

The kernel function must be:

- symmetric [$K(x_i, x_j) = K(x_j, x_i)$]
- positive semi-definite: if $x_1, \ldots, x_n \in S$, then the $n \times n$ matrix $G$ (Gram matrix or kernel matrix) defined by $G_{ij} = K(x_i, x_j)$ is positive semi-definite*

* has non-negative eigenvalues
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 $f(x)$ we can just replace the dot products with kernels $K(x_i,x_j)$:

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

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

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

- **Linear regression:**

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

- **Class membership:**

\[
p(y = true|x) = \sum_{i=0}^{N} w_i \times f_i = w \cdot f \quad (3)
\]
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: $y_1, y_2, \ldots, y_C$
- We estimate the probability that $y$ is a particular class $y$ given $N$ feature functions as:

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

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

Markov models

Directional Models

Logistic Regression

Linear-chain CRF

CRF

Adapted from C. Sutton, A. McCallum, “An Introduction to Conditional Random Fields”, ArXiv, November 2010
• **Linear chain conditional random field:**
  
  – Let \( X = (x_1, \ldots, x_T) \) be a random variable over data sequences to be labeled and \( Y \) a random variable over the corresponding label sequences
  
  – All components \( y_j \) of \( Y \) are assumed to range over a finite label alphabet \( \sum \)
  
  – 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
• 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]
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

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

$\lambda_i = \text{parameter that models the observed statistics in the training examples}$

$Z = \text{normalizing constant}$

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

$$Y^* = \arg\max_Y p(Y|X)$$
• **CRF training:**
  
  – Like for the Maxent model, we need numerical methods in order to derive $\lambda_i$
  
  – 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 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)
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]
For these reasons, the Commission by a majority declares the application admissible, without prejudging the merits.

It follows that the application cannot be dismissed as manifestly ill-founded.

The Commission has taken cognizance of the submissions of the parties.

In these circumstances, the Commission finds that the application cannot be declared inadmissible for non-exhaustion of domestic remedies.

The Commission recalls that article 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.

The Commission notes that in the context of the section powers the secretary of state has a very wide discretion.

The Commission recalls that in the case of temple v. the united kingdom no. x dec. d.r. 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)

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

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

Figure 1. (a) The spatial ontology. (b) Example sentence and the recognized spatial concepts.
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
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 $W_i$
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 (\( \Psi \)) specifications.

Can be written in terms of the instantiations of the templates and their related blocks of weights \( W_p \)
A popular discriminative training approach is to minimize the following convex upper bound of the loss function over the $N$ training data:

\[
\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\
\text{s.t. } \forall i, \forall y \in \mathcal{Y} \setminus y_i : \quad \mathbf{w}^T \Psi(x_i, y_i) \geq \mathbf{w}^T \Psi(x_i, y) + \Delta(y_i, y) - \xi_i
\]

- Training with the most violated constraints/outputs ($y$) per training example
- In the experiments: structured support vector machines (SSVM), structured perceptrons and averaged structured perceptrons
Constraints are linear and variables take the form of integers.

Constraints are applied:
- during training: finding the most violated outputs
- and/or
- during testing
E.g., Markov random fields

Allow using rules as features for which the weight is trained on the annotated data

Concern: the computational complexity
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]
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, ...
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
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

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

Conclusion

Premises
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

Human understanding of text: humans connect to their world/domin 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: 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.
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/
TE provides techniques to detect both the argument components, and the kind of relation underlying them:
Or an entailment or a contradiction is detected
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

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>
<thead>
<tr>
<th>Classifier Combination</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max.Ent. and Support Vector Machine</td>
<td>77.49</td>
<td>60.88</td>
<td>74.07</td>
</tr>
<tr>
<td>Context-free Grammar</td>
<td>61.00</td>
<td>75.00</td>
<td>67.27</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Classifier Combination</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max.Ent. and Support Vector Machine</td>
<td>70.19</td>
<td>66.16</td>
<td>68.12</td>
</tr>
<tr>
<td>Context-free Grammar</td>
<td>59.00</td>
<td>71.00</td>
<td>64.03</td>
</tr>
</tbody>
</table>

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

[Mochales & Moens AI & Law 2011]
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

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

Table 1: Example Sentences
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

Table 4: # of propositions in Train and Test Set

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>UNVERIF vs All</th>
<th>VERIFNON vs All</th>
<th>VERIFEXP vs All</th>
<th>Average F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNI(base)</td>
<td>85.24</td>
<td>79.43</td>
<td>82.23</td>
<td>42.57*</td>
</tr>
<tr>
<td>UNI+Bi</td>
<td>82.14</td>
<td>89.69*</td>
<td>85.75*</td>
<td>51.67*</td>
</tr>
<tr>
<td>VER</td>
<td>88.52*</td>
<td>52.10</td>
<td>65.60</td>
<td>28.41</td>
</tr>
<tr>
<td>EXP</td>
<td>82.42</td>
<td>4.45</td>
<td>8.44</td>
<td>20.92</td>
</tr>
<tr>
<td>VER+EXP</td>
<td>89.40*</td>
<td>49.50</td>
<td>63.72</td>
<td>29.25</td>
</tr>
<tr>
<td>UNI+Bi+</td>
<td>86.86*</td>
<td>83.05*</td>
<td>84.91*</td>
<td>49.88*</td>
</tr>
</tbody>
</table>

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

[Park & Cardie FWAM 2014]
RECOGNIZING ARGUMENTS IN ONLINE DISCUSSIONS

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

<table>
<thead>
<tr>
<th>Label</th>
<th>Description: Comment...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>...explicitly attacks the argument</td>
</tr>
<tr>
<td>a</td>
<td>...vaguely/implicitly attacks the argument</td>
</tr>
<tr>
<td>N</td>
<td>...makes no use of the argument</td>
</tr>
<tr>
<td>s</td>
<td>...vaguely/implicitly supports the argument</td>
</tr>
<tr>
<td>S</td>
<td>...explicitly supports the argument</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Labels</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>a</td>
<td>N</td>
<td>s</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>UGIP</td>
<td>48</td>
<td>86</td>
<td>691</td>
<td>58</td>
<td>130</td>
<td>1,013</td>
</tr>
<tr>
<td>GM</td>
<td>89</td>
<td>73</td>
<td>849</td>
<td>98</td>
<td>176</td>
<td>1,285</td>
</tr>
<tr>
<td>UGIP+GM</td>
<td>137</td>
<td>159</td>
<td>1,540</td>
<td>156</td>
<td>306</td>
<td>2,298</td>
</tr>
</tbody>
</table>

Table 5: Distribution of labels in the COMARG corpus
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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>A-a-N-s-S</th>
<th>Aa-N-sS</th>
<th>A-N-S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UGIP</td>
<td>GM</td>
<td>UGIP</td>
</tr>
<tr>
<td>MCC baseline</td>
<td>68.2</td>
<td>69.4</td>
<td>68.2</td>
</tr>
<tr>
<td>BoWO baseline</td>
<td>68.2</td>
<td>69.4</td>
<td>67.8</td>
</tr>
<tr>
<td>TE</td>
<td>69.1</td>
<td>81.1</td>
<td>69.6</td>
</tr>
<tr>
<td>STS</td>
<td>67.8</td>
<td>68.7</td>
<td>67.3</td>
</tr>
<tr>
<td>SA</td>
<td>68.2</td>
<td>69.4</td>
<td>68.2</td>
</tr>
<tr>
<td>STS+SA</td>
<td>68.2</td>
<td>69.5</td>
<td>67.5</td>
</tr>
<tr>
<td>TE+SA</td>
<td>68.9</td>
<td>72.4</td>
<td>71.0</td>
</tr>
<tr>
<td>TE+STS+SA</td>
<td>70.5</td>
<td>72.5</td>
<td>68.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>UGIP → GM</th>
<th>GM → UGIP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A-a-N-s-S</td>
<td>Aa-N-sS</td>
</tr>
<tr>
<td>STS+SA</td>
<td>69.4</td>
<td>69.4</td>
</tr>
<tr>
<td>TE+SA</td>
<td>72.6</td>
<td>73.5</td>
</tr>
<tr>
<td>STS+TE+SA</td>
<td>71.5</td>
<td>72.2</td>
</tr>
</tbody>
</table>

Boltužić & Šnajder FWAM 2014
Opinion mining: finding arguments and counter arguments for an opining 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]
Mining of the supporting evidence of claims in scientific publications and patents and their visualization for easy access

Scientific idea + Expectations + Observations = Scientific argument

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