Language Technologies for Information Services

Marie-Francine Moens
HMDB-LIIR
Department of Computer Science
Katholieke Universiteit Leuven, Belgium
Problem

- We have many information sources:
  - Course documents, publications, emails, blogs, discussion fora, pictures, slides, speech, graphics, ....
  - How can **language technology** help us using them, relating them together, acquiring knowledge from it?
Language technologies

- Text categorization
- Topic modeling
- Semantic frame recognition
- Named entity recognition and noun phrase coreference resolution
- Rhetorical classification
- Language reformulation
- Question answering
- ...
Text categorization

- assigning controlled language descriptors to texts
- Usually categorization of a complete document
- Descriptors or labels
  - A priori defined
  - Concept terms, subject or classification codes: $C = \{C_1, C_2, \ldots, C_m\}$
  - Labels might be organized in hierarchy
Business Schools Sprout Greener MBAs

Sunday September 23, 1:57 pm ET
By Michelle Locke, Associated Press Writer

Green MBAs Seek to Balance Profit and Planet

OAKLAND, Calif. (AP) -- Business professor John Stayton remembers when eyes would start rolling at the idea of a "green MBA."

These days, business schools across the country are incorporating the environmental and social costs of doing business into their curricula, and a few, like the program Stayton directs at Dominican University of California, aim for an all-green program.

The goal? How to succeed in business without really frying the planet.

"Essentially we've got to change the way we're doing everything and making everything," said Stayton.

The program Stayton directs was launched at Santa Rosa's New College of California North Bay in 2000 as a Master of Arts in the humanities department and transferred to Dominican last spring. It's one of a handful of such degrees; others include MBAs offered at the Presidio School of Management in San Francisco and the Bainbridge Graduate Institute in Washington state.

The move to balance economy and ecology is showing up all over, said Rich Leimsider, director of the Center for Business Education at The Aspen Institute, a leadership think tank which reports on how Master of Business Administration programs are adding social and environmental issues to their courses in its biennial "Beyond Grey Pinstripes" report.

"It matters what the senior executives of companies do, say and think," said Leimsider. "If you can change business education to include an appreciation for the social and environmental context you wind up with leaders who are really good at creating..."
Sharapova beats Ivanovic to win Australian Open

A year after being on the wrong end of one of Russian didn’t drop a set in seven matches at Melbourne Park, including wins the most-lopsided losses in a Grand Slam final, Sharapova wrapped up her third major title with a 7-5, 6-3 victory over fourth-seeded Ana Ivanovic on Saturday. The 20-year-old over three of the top four ranked players, erasing 12 months worth of painful memories in the wake of her 6-1, 6-2 loss to Serena Williams last year. After Ivanovic sprayed a forehand wide on match point, Sharapova dropped to her knees and appeared to be fighting back tears as she waved and blew kisses to the crowd. Then she dropped her racket in her chair before heading to shake hands and exchange high-fives with her father and supporters.

Great victory!
Methods

- Supervised machine learning:
  - Train classifier with examples manually annotated with classes (support vector machine, logistic regression, multinomial Bayes, ...)
  - Apply the learned model to unseen examples
- Focus on feature extraction: chi-square, information gain, linear classifier weights, frequent item sets, statistical dimensionality reduction, ...
- Focus on reduction of supervision: active learning

EU FP6-027978
Anticipatory Learning for Reliable Phishing Prevention

KRO/009/06
TIME
IWOIB Attentio

Acknowledgements
IBBT
Potential applications e-learning

- Subject categorization of content
- Automated generation of content metadata
Latent semantic topic models

= a class of unsupervised (or semi-supervised) models in which the semantic properties of words and documents are expressed in terms of topics

- models are also called aspect models

Latent Semantic Indexing:
- the semantic information can be derived from a word-document matrix

[Deerweester et al. 1990]

But, LSI is unable to capture multiple senses of a word

Probabilistic topic models
Probabilistic topic model

- **Generative model** for documents: probabilistic model by which documents can be generated
  - document = probability distribution over topics
  - topic = probability distribution over words

To make a new document, one chooses a distribution over topics, for each topic one draws words according to a certain distribution:

- select a document $d_j$ with probability $P(d_j)$
- pick a **latent class** $z_k$ with probability $P(z_k|d_j)$
- generate a word $w_i$ with probability $P(w_i|z_k)$

[Steyers & Griffiths 2007]
observed word distributions

word distributions per topic

topic distributions per document
Probabilistic Latent Semantic Analysis (pLSA)

John goes into the building, sits down, the waitress shows him the menu. John orders. The waitress brings the food. John eats quickly, puts $10 on the table and leaves. ...

John goes to the park with the magnolia trees and meets his friend, ...

[Hofmann SIGIR 1999]

\[ M = \text{number of documents} \]
\[ N = \text{number of words} \]
Methods

- Latent semantic indexing
- Probabilistic latent semantic analysis
- Latent Dirichlet Allocation [Blei et al. JMLR 2003]

Class
EU FP6-027978

WebInsight

AMASS++
IWT (SBO 060051)
Potential applications in e-learning

- Linking content: when sufficient overlap in topic distributions, probabilistic modeling of similarities and differences in content, ...
- Linking content across languages
Semantic frame labeling

- Recognizing the basic event structure ("who" "does what" "to whom/what" "when" "where" ...): semantic roles that form a semantic frame => determining the meaning of a sentence

- Important task for natural language understanding: question answering, information extraction
Recognizing the basic event structure of a sentence ("who" “does what” “to whom/what” “when” “where” ...):

Maria Sharapova walks towards the field.

Maria Sharapova kisses the winner's trophy at Melbourne Park
Methods

- Supervised learning methods:
  - Features linguistically motivated
- Focus on semi-supervised models

EU FP6-027978

IWT (SBO 060051)
Potential applications in e-learning

- Basic understanding of sentences, captions of images, ...
- Relationship detection between entities
Named entity recognition

- recognizing the semantic class of (proper) names: e.g., person, location, organization

Noun phrase coreference resolution

- identifying an equivalence relation between noun phrases (nouns, pronouns) in the discourse
planemakers Boeing and Airbus vie for Qantas contract

December 4, 2005: 12:19 AM EST

SYDNEY (Reuters) - Qantas Airways Ltd., which will seek board approval this week to spend up to A$20 billion ($15 billion) on new planes, said on Sunday the contest between rival manufacturers was the closest it has seen.

Planemakers Boeing and Airbus are set to end a record year for new plane orders with a decision from Australia's Qantas, which has said it might need as many as 100 new planes.

Qantas chief executive Geoff Dixon said management had not decided on a final recommendation for Wednesday's board meeting, but fleet renewal was essential for the carrier, which also wants to expand its low-cost Jetstar airline internationally …
Clinton was the third-youngest president, behind Theodore Roosevelt (the youngest president) and John F. Kennedy (the youngest to be elected president). He was the first baby boomer president. ...

The husband of Hillary Clinton ...
Methods

- Supervised learning (see supra)
- Noun phrase coreference resolution:
  - Learning of pairwise relations, followed by clustering of the mentions
  - Linguistically oriented features
Potential applications in e-learning

- Detection and resolution of persons, organizations, dates, locations, domain specific entity names such as protein names, animal species, ...
- Assisting technologies for other tasks
Rhetorical segmentation and classification

- Segmenting the discourse into elements
- Recognizing the rhetorical relationships between the elements
- Many different taxonomies of rhetorical relations
- Examples of rhetorical relations: illustration, contrast, elaboration, summary, premise and conclusion in argument, ...
Checked baggage allowance

Extra security measures for hand baggage are in place at all airports.

The amount of baggage that may be checked onto flights operated by KLM, Air France, Northwest Airlines and Kenya Airways free of charge, otherwise known as the baggage allowance, is stated below.

If your flight is operated by another KLM partner airline, enquire with your travel agent whether any other baggage restrictions apply.

Please bear in mind that certain items are forbidden.

On this page:

- Checked baggage
- Baggage for children under two years of age
- Arrange extra baggage online and get a 30% discount!
- Exceptions to baggage allowance
- Baggage to be transported as dangerous goods
- KLM Partner Airlines

Checked baggage

The amount of baggage that you can take with you free of charge is determined according to either the number of pieces or the weight. The ‘allow’ box on your paper ticket indicates which of these concepts apply to your baggage allowance. On your e-ticket, this is indicated in the “baggage” box.
{ [SUPPORT]: The Court recalls that the rule of exhaustion of domestic remedies referred to in Article x of the Convention art. x obliges those seeking to bring their case against the State before an international judicial or arbitral organ to use first the remedies provided by the national legal system.  

CONCLUSION: Consequently, States are dispensed from answering before an international body for their acts before they have had an opportunity to put matters right through their own legal systems. }

[SUPPORT]: The Court considers that, even if it were accepted that the applicant made no complaint to the public prosecutor of ill-treatment in police custody, the injuries he had sustained must have been clearly visible during their meeting.  

AGAINST: However, the prosecutor chose to make no enquiry as to the nature, extent and cause of these injuries, despite the fact that in Turkish law he was under a duty to investigate see paragraph above.  

SUPPORT: It must be recalled that this omission on the part of the prosecutor took place after Mr Aksoy had been detained in police custody for at least fourteen days without access to legal or medical assistance or support.  

SUPPORT: During this time he had sustained severe injuries requiring hospital treatment see paragraph above.  

CONCLUSION: These circumstances alone would have given him cause to feel vulnerable, powerless and apprehensive of the representatives of the State.  

CONCLUSION: The Court therefore concludes that there existed special circumstances which absolved the applicant from his obligation to exhaust domestic remedies.  

Figure 2: Example of legal argumentation
Methods

- Predictable structures: e.g. argument tree: context free grammar, induction of the grammar
- Focus on learning with few examples

ACILA
OT/06/03

DAISY
Stevin ST-07-015
Potential applications in e-learning

- Content generation: reusable components
Sentence rephrasing

- reformulating a sentence:
  - Paraphrasing
  - Compression
“SNS Bank heeft maatregelen getroffen voor veilig Internet Bankieren”
(SNS Bank has taken measures to perform bank transactions in a safe way).

In the context of the discourse, the sentence is reduced to

“Maatregelen voor veilig Internet Bankieren”
(Measures to perform bank transactions in a safe way).
(CNN) -- Russian energy giant Gazprom, in the midst of a standoff with Ukraine over payments for gas, has filed a lawsuit to stop what it claims is illegal siphoning.

(CNN) - Gazprom is a Russian energy giant. Russia has a dispute with Ukraine over payments for gas. Russia has filed a lawsuit. It claims illegal withdrawing of gas.

A gas-compressor and gas-holder station in Mryn, Ukraine. Gazprom has stopped supplying the country.

A gas-compressor and gas-holder station. Gazprom has stopped supplying the country, Ukraine.
Methods

- Basics: word sense disambiguation, named entity recognition
- Rule based syntax reformulations
- Learning of valid compressions or paraphrases (e.g. noisy channel models found in machine translation)
- Heuristic rules for selection of valid paraphrase
Potential applications in e-learning

- Reformulation of content to adapt to users
- Matching of similar information
Question answering

- **Automatic question answering:**
  - single questions are automatically answered by using a collection of documents as the source of data for the production of the answer:

- Many different types of questions:
  - Factual questions
  - List questions
  - Definition questions
  - Natural language statement
  - ...
Example:

**question:** “Who is the architect of the Hancock building in Boston?”

**answer:** “I.M. Pei”

**extracted from:**
“The John Hancock Tower was completed in 1976 to create additional office space for the John Hancock Life Insurance Co. It was designed by the renowned architect I.M. Pei.”

“Designed by world renowned architect I.M. Pei, the John Hancock Tower is the highest in New England.”
Example:

**Natural language query:** “Show me a video fragment where a red car takes a right turn on Saint-John’s square.”

**answer:**
keyframes of video fragment

**extracted from:**
video indexed with entities their attributes and relations including spatial and temporal relations
Methods

- Syntactic and semantic analysis of the question and potential answers in the documents
- (Relaxed) matching of the meaning representations
Potential applications in e-learning

- Search!
Questions?

www.cs.kuleuven.be/~liir/