

Machine Learning and Open Courseware



Outline

- What is **open courseware**?
- Is there still research to be done?
- Recommendation
- Translation and Transcription
- Annotation
- Navigation
- Traces

Opencourseware=Open Educational Resources

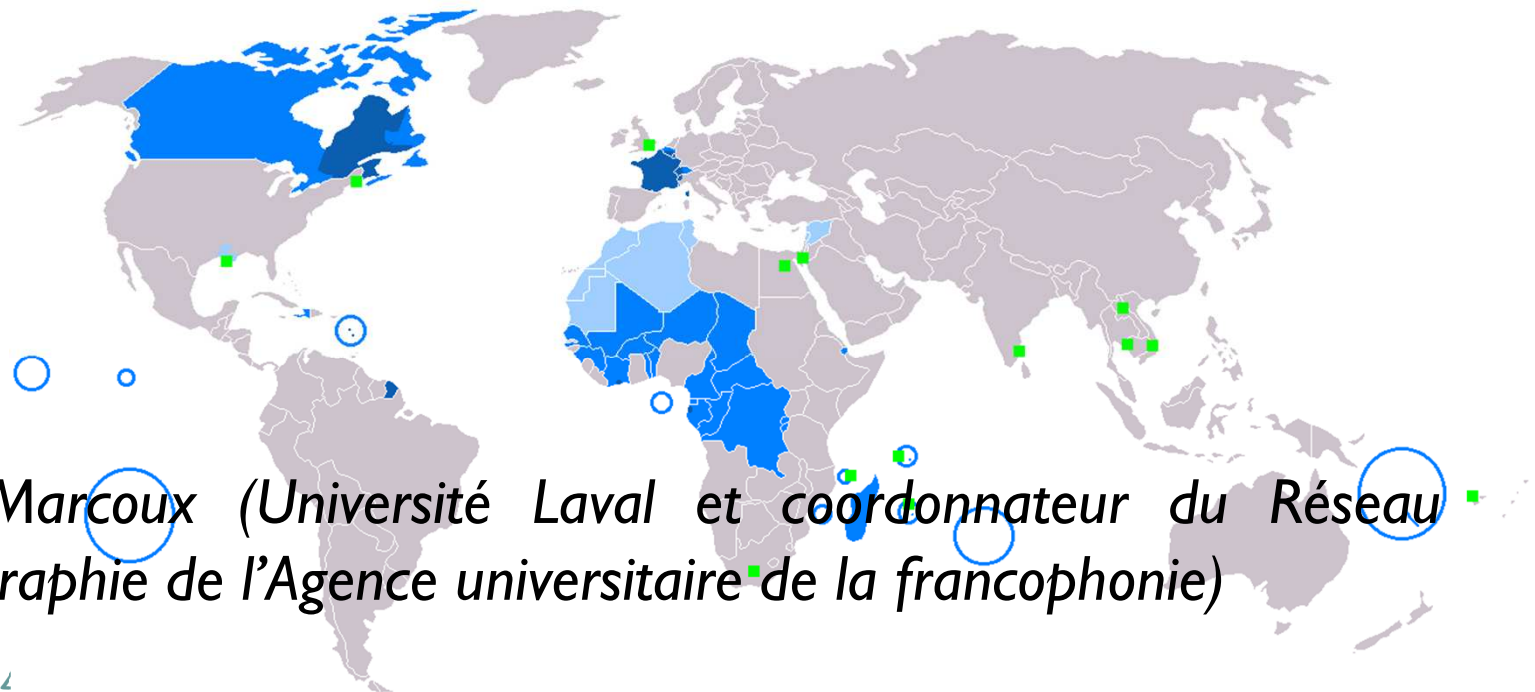
- Organization: OCWC
 - 30 000+ Modules
 - 280+ Organizations
 - 40 Countries
 - 29 Languages

Some reasons (I)

- *By 2025, the global demand for higher education will **double** to ~200M per year, mostly from emerging economies (NAFSA 2010)*
- 100 M/15 years... **80 000 new students/week!**
- Diana Laurillard.
<http://www.ucd.ie/teaching/u21conference2013/>
- Anka Mulder <http://ankamulder weblog.tudelft.nl/>

Some reasons (2)

- 220 M French Speakers in 2010... 700 M in 2050
- 3% (2012) → more than 7% in 2050
- By 2030, more French Speakers than English Speakers (only 5% in 2050)



- René Marcoux (Université Laval et coordonnateur du Réseau Démographique de l'Agence universitaire de la francophonie)

How does this relate to MOOCs?

- Historically, appeared before MOOCs
- Most MOOC players acknowledge inheritance
- Also, through MOOCs only the upper layer of what we teach is being shared
- Question is: can we do something with the courseware that has been produced so far?
- Some believe we can do better

OpenCourseware and Pascal

- The Pascal network ran 2003-2012
- One key asset was videolectures.net
- PASCAL 2 → Knowledge 4 All foundation
- Goal is to build upon the Pascal legacy



VideoLectures current stats

Content

Events : 800
Authors: 11,000
Lectures: 15,000
Videos: 17,000
Organisations: 7,000
Categories : 600
Comments: 8500

Website

Video views: 6 million
Page views: 27 million
Signed in users:
25,000
Average time: 36 min
Attachments: 620,000
Files: 1,3 million
8 servers
New Visitor: 60.83%
Licenses: Creative
Commons

Users

Top countries: United
States, India, Slovenia,
UK, Germany, China,
Canada

Tutorials: 350
Keynote: 800
Interviews: 250

32  2,525,076



Cdlh 2014

<http://conference.ocwconsortium.org/2014/>

OCWC
GLOBAL
CONFERENCE

LJUBLJANA, SLOVENIA
23 - 25 April 2014

Open Education for a Multicultural World

ABOUT | CALL FOR PAPERS | SCHEDULE | REGISTRATION | VENUE | PARTNERS | PAST CONFERENCES

CONTACT US



Cdlh 2014

lina LABORATOIRE D'INFORMATIQUE
DE NANTES ATLANTIQUE

ARE THERE REALLY SOME RESEARCH QUESTIONS THERE?



Cdlh 2014

Research is e-learning?

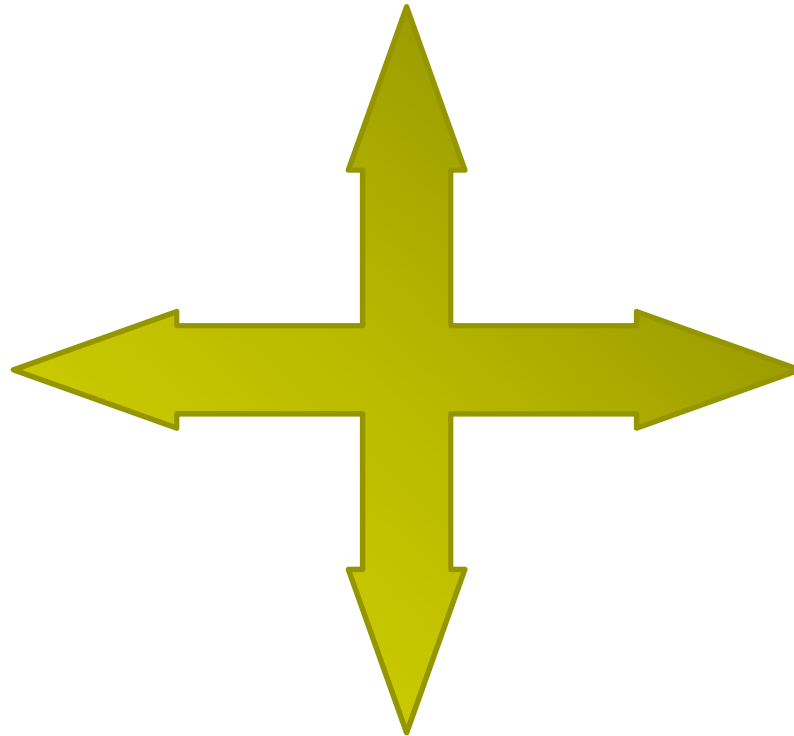
- In part. Lots of new scenarios
- Much more learning data which needs analysing

- But not only

A simplified picture

The learner and his experience (**e-learning**)

Sharing open
educational
resources



Knowledge as a
merchandise

Helping the teacher deal with the masses of
information (**machine learning**)

Clue: Think about MOOCs

- Coursera: Andrew Ng and Daphne Koller
- Eduacity,
- Keywords: automatic evaluation, learning analytics, big data

Clue: Some titles of talks from the Internet of Education conference, Slovenia, November 2013

http://www.k4all.org/Internet_of_Education/

- TransLectures: cost-effective transcription and translation of video lectures
- Knowledge Building Through Collaborative Hypervideo Creation
- Annotations, a key asset for video-based e-learning
- A MediaMixer for online learning? – making learning materials more valuable for their owner and more useful for their consumer

Clue: OCWC 2014 Global

- Stats from the accepted papers
- 4 Tracks

Track	# papers	# posters
Open Educational Policies	16	
Research and Technology	20	13
Pedagogical Impact	20	4
Project dissemination	10	

Slides based on those by Matjaz Rihtar <matjaz.rihtar@ijs.si>

MACHINE LEARNING AND RECOMMENDATION



Cdlh 2014

Why recommendation?

- Simplest question is: after this video, which one should I consider?
- Basic system uses history of transactions: *most viewers who watched A also watched B*
- Can we use?
 - Individual history
 - Thematic similarity
 - History over what the users really do

Machine learning approach

- Extract many different features
- Let the learning algorithm put different weights on these depending on their usefulness
- Think carefully about the validation issues in order to verify what you have learnt

Obviously a slide improper for SML

Experience from *LaVie*

- Were used for recommendation
 - Transcriptions
 - Graphs of authors
 - Related pdfs
 - Other ontologies (Wikipedia, DBLP and Google)

Topic and user modeling

- **7 features:**

1. Lecture popularity
 - Number of visits
2. Content similarity
 - $\text{BoW}(L_c) \cdot \text{BoW}(L_p)$
3. Category similarity
 - $\text{BoC}(L_c) \cdot \text{BoC}(L_p)$
4. User content similarity (**computed on the fly**)
 - $\text{BoW}(\text{Hist}(U)) \cdot \text{BoW}(L_p)$
5. User category similarity (**computed on the fly**)
 - $\text{BoC}(\text{Hist}(U)) \cdot \text{BoC}(L_p)$
6. Co-visits
 - Number of times of L_c and L_p viewed in the same browsing session
7. User similarity
 - Number of users who have watched both L_c and L_p

L_c = current lecture

L_p = proposed lecture

U = user

I table has approx. 70 million entries
(for features 2,3,6 and 7)

Recommendation (I)

- Using SVM classifier for training:
 - Positive samples: two months of clicks using current recommender
 - Resulting feature weights:

Feature	Weight
Lecture popularity	-0.00003
Content similarity	0.00452
Category similarity	0.00148
User content similarity	0.02724
User category similarity	0.04167
Co-visits	0.00187
User similarity	0.01519

Recommendation (2)

- Final recommendation
 - A linear SVM classifier was used to rank all possible recommendation links:

Given L_c and U :

For all $L_p \neq L_c$:

\vec{x} ... feature vector computed for the triplet (L_c, L_p, U)

$$\text{score}(\vec{x}) = \vec{w} \cdot \vec{x} = \sum_{n=1}^7 w_n \cdot x_n$$

- Lectures with top 10 scores are recommended

Evaluation

- Evaluation
 - Using coin flipping between old and new recommender
 - Counting the number of clicks
- Try <http://dev.videolectures.net/>

The screenshot shows a video lecture page for "Road safety in Europe". The main content area includes the title, presenter information (Jerry Hole, Rijkswaterstaat), publication and recording dates (May 22, 2008, April 2008), and view count (45). It also features a category breadcrumb (Top » Business » Transportation and Logistics » Traffic Safety), a popularity rating (5 stars), and social sharing options (Tweet, Me gusta, +1, Share). Below this is a "Link this page" section with a form to request a link to the lecture on a homepage, and a "Write your own review or comment" section with input fields for Name, Email address, URL, and Comment.

On the right side, there is a sidebar titled "Visitors who watched this lecture also watched...". It lists several related video lectures with their titles, view counts, and presenters:

- Traffic safety (69 views - Jean Yves Le Coz, 2008)
- Best practices for road safety in Europe: A systematic approach (99 views - Martin Winkelbauer, 2008)
- An integrated approach to PTW road safety (53 views - Federico Galliano, 2008)
- Road safety and its modification (97 views - Ljubo Zajc, 2008)
- Using micro-stimulation modelling for driver assistance system assessment (217 views - Evangelia Gaitanidou, 2008)
- Road Safety Policy (36 views - Jean Yves Le Coz, 2008)
- Debate - Traffic Safety (52 views - Jean Yves Le Coz, Ken Ducatel, Herman Meyer, Jean Lalo, Jerry Hole, 2008)
- Active Safety (78 views - Mathias Schulze, 2008)
- Razprava o kontinuiteti in novostih v 7. OP (37 views - 2007)
- Intelligent car = safer roads (98 views - Ken Ducatel, 2008)



TRANSCRIPTION AND TRANSLATION



Cdlh 2014

transLectures



— Transcription and Translation of Video Lectures

Slides by Gonçal Garcés

Project coordinator: Alfons Juan-Ciscar

ggarces@dsic.upv.es

ajuan@dsic.upv.es



The transLectures partners

	Name	Country
1	Universitat Politècnica de València	Spain
2	Xerox SAS	France
3	Institut Jožef Stefan	Slovenia
3+	Knowledge for All Foundation	UK
4	RWTH Aachen University	Germany
5	EML – European Media Laboratory	Germany
6	DDS – Deluxe Digital Studios	UK

The transLectures approach

1. Automatic Speech Recognition (ASR) and Machine Translation (MT)
 - Adaptation: Taking advantage of the characteristics of video lecture repositories
 - High-quality automatic transcriptions and translations
2. Interactive postediting: intelligent interaction for reduced effort

Goals

- Massive adaptation
- Intelligent interaction

- Implementation
 - Case studies: Videolectures.NET & Polimedia
 - Real-life evaluation
- Integration into Opencast Matterhorn

<http://opencast.org/matterhorn/>



Languages

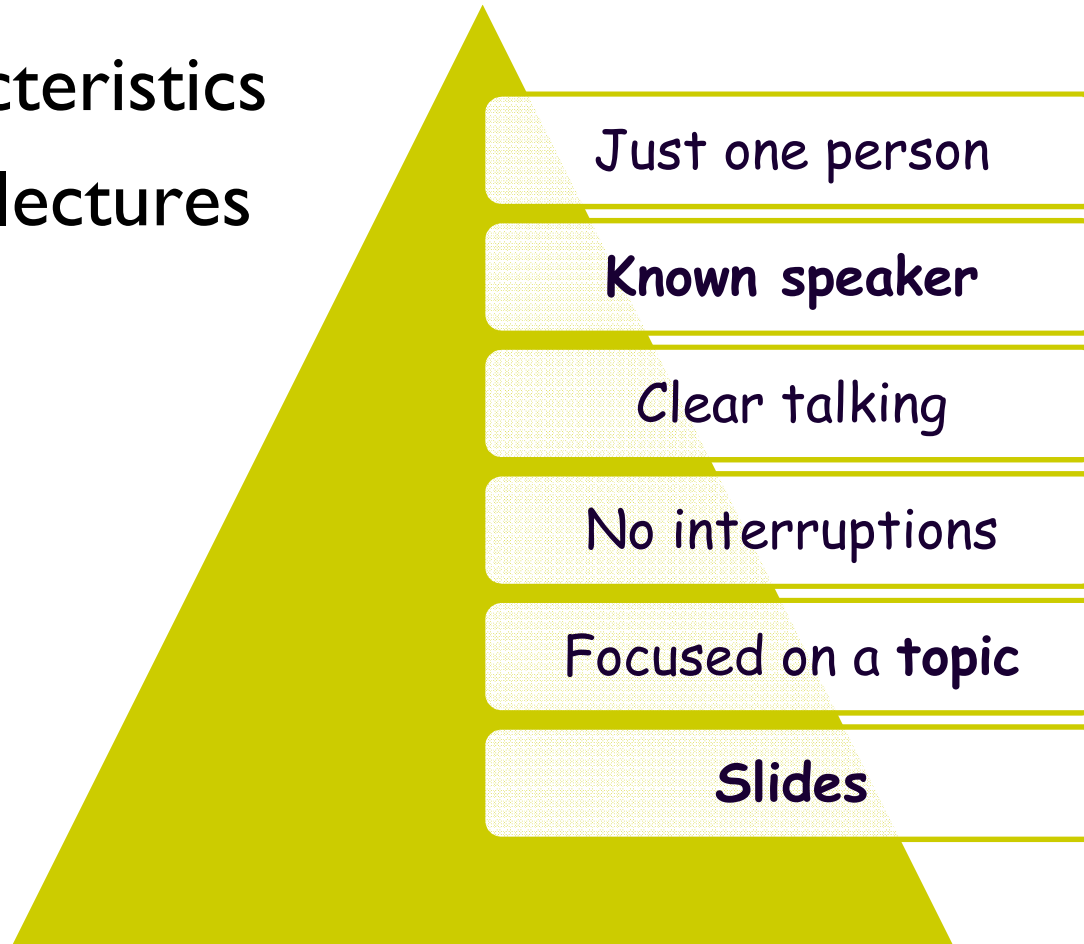
- Transcription (ASR)
 - EN
 - SL
 - ES
- Translation (MT)
 - EN>SL , SL>EN
 - EN>ES , ES>EN
 - EN>FR
 - EN>DE

transLectures: video demo



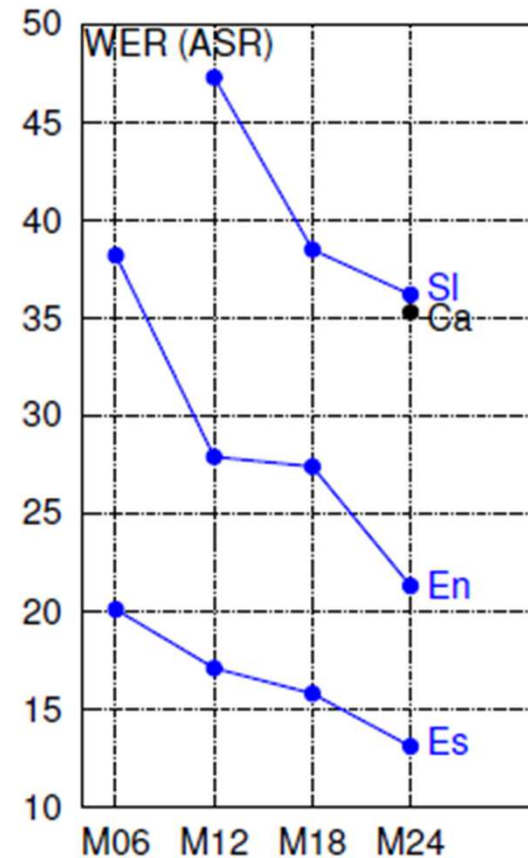
Massive adaptation

- Characteristics of video lectures



Scientific evaluations

- Transcription results
 - WER: Word Error Rate (%)
 - Goal: WER < 25%
 - EN, SL, ES



Worse

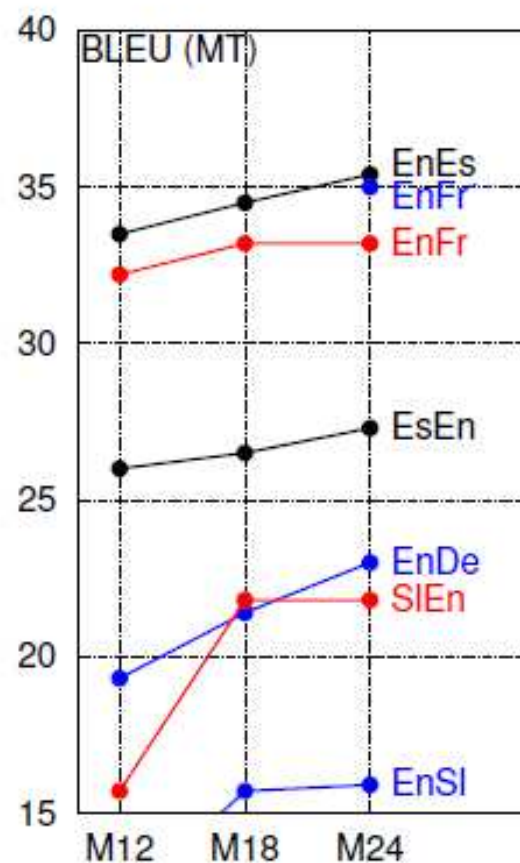
Better

Scientific evaluations

- Translation results

- BLEU
- Goal: BLEU > 30

- EN>SL , SL>EN
- EN>ES , ES>EN
- EN>FR
- EN>DE



Better

Worse

YI results and comparison

- Videolectures.NET:

WER	tL	Google
English	27.9	45.4
Slovenian	44.4	N/A

- poliMedia:

WER	tL	Google
Spanish	17.1	34.3

YI results and comparison

- Videolectures.NET:

			BLEU	
WER	tL	Google	tL	Google
English	27.9	45.4	English → Spanish	33.5 34.1
Slovenian	44.4	N/A	English → French	32.2 32.0
			English → German	20.6 18.6
			English → Slovenian	15.5 10.5
			Slovenian → English	15.7 17.5

- poliMedia:

			BLEU	
WER	tL	Google	tL	Google
Spanish	17.1	34.3	Spanish → English	26.0 27.6

Intelligent interaction

- Post-editing automatic transcriptions/translations
 - The user invests the least possible effort
 - The system **learns** the most from it
- Confidence measures
- Fast constrained search

Intelligent interaction

Selected word before supervision

hemos **hace** equivalencia al **derecho** internacional y contra ponerlo al propio derecho interno de los estados **para ver** las características que le diferencian //

[sonido de fondo]

en concreto derecho internacional público se caracteriza por unas

Selected word after supervision

hemos **hecho** equivalencia al **derecho** internacional y contra ponerlo al propio derecho interno de los estados **para ver** las características que le diferencian //

[sonido de fondo]

en concreto derecho internacional público se caracteriza por unas

User evaluations

2. Intelligent interaction

nos puede interesar o no pero podemos determinar el tiempo en que sabemos que

aunque sea de nuestro interés sabremos determinarlo consta de las historias clínicas el tiempo

consta en los registros del tiempo que ocurren dos veces

 porque si estamos interesados y no « consta » es como

e imposible de plantear un estudio de tipo de incidencia

esto vamos a estudiar por lo tanto incidencias que sería interesados por el tiempo que se puede determinar

o prevalencia esto viene incidencias acumuladas

en el fondo viene a ser muy similar

dado que una prevalencia no es más que el resultado

de una incidencia acumulada en un margen de tiempo determinado

User evaluations

- User evaluations at UPV: results

Table 2: Comparison between all the interaction models evaluated.

Supervision mode	Initial WER	Final WER	Lecturer RTF	SS-score
1st - Manual supervision	16.9	0.0	5.6	9.1
2nd - Intelligent interaction	14.5	8.0	2.2	7.2
3rd - Two-step supervision	28.4	0.0	5.3	7.8

transLectures: Open source tools

- The tL player (& editor)
 - Coming soon (www.translectures.eu)
- The transLectures-UPV Toolkit (TLK) for ASR
 - www.translectures.eu/tlk
- RWTH Aachen: rASR, Jane (MT)
 - <http://www-i6.informatik.rwth-aachen.de/web/Software/>

Multilingualism

- Multilingualism is a crucial issue
 - If English is the unique language we have many deprived users
 - But also we are missing a lot of excellent material
- It is a highly sensitive issue
- Answers
 - Being able to navigate between languages
 - Being able to translate: translectures project

Olivier Aubert - @Olivier_Aubert
Yannick Prié - @yprie

ANNOTATION ISSUES



Cdlh 2014

Key questions

- Obtaining the participation of the learner can be vital (at least if the main purpose of the resource is to allow a learner to learn)
- Examining the MOOC's success helps
- Why would someone want to annotate our videos?
- What tools can we provide in order for quality annotation to take place ?

Video-based e-learning activities

- Different activities based on
 - the nature of the video document
 - the status of the annotator
 - the status of the recipient

Assimilation - example

Advène - MZS_yuki.azp - Cours d'expérience, Yuki, Musée zoologique, Strasbourg, avril 2011

File Edit View Player Packages Help

No active dynamic view

Popups Verbalisatio... x

V: mh I think it's a little bit bored it's not + it's just really a kind of collection for me + it's like there are collections + we have this kind of collections of birds this: huge amount + it's really huge amount <C: mh mh> but we didn't really show out like + were they live <C: mh> and where like + where they come from + what kind of environment do they live <C: mh> and also for this kind of the-the little notes it's not really clear because like + at least when you write where do get this bird this-this exact one and what's the species name in latin and maybe in english or in germany I don't know but generally in french but it's not really that good for-to people who doesn't really-doesn't know anything about zoo knowledge to understand what they are <C: mh> they are just collections like-for a professor of animal behaviour or animal or zoo knowledge + they will understand but for no more poeple it's a little bit bored
 C: so you just pass
 V: I just pass yea I couldn't even-I couldn't remember all the birds I cant
 C: so sometime you look at d'vous we looks at one of w

00:08:43.743

Timeline x

Scale 6.570|00:07:56.080|00:08:55.590|00:09:55.100|00:10:54.610|00:11:54.120|00:12:53.630|00:13:53.140|00:

Discrete scrolling

42%

Verbalisations V: I know it | C: V: mh I think it's a | V: QV: ok so | V: oh I say- V: I t | V: eve | V: n | C: V: l'm | V: ye | V: | V: sl | C: V: l th | V: V: l d | V:

Représentamen R La gale La gale Les oiseaux et es Les mâles Les

Engagement E Éviter Essayer Essayer de trouv Essayer de trouver des Env

Anticipation A Attent Attentes liées à c Attentes liées à la prése Atte

Référentiel S Ce mus Signe hexadique n°3 Chez certa Nou

Unité de cours d'expérien Essayer de comprendre les intentions du musée sans y arriver Cor

Interprétat L

V: mh I think it's a little bit bored it's not + it's just really a kind ... I understand but for no more poeple it's a little bit bored (a55) 00:08:27.147 - 00:09:27.309

Advène: an example of an offline anotator

Collaborative assimilation example

Firefox

Boîte de réception ... 0 messages non lus... Communication - ... Annotations x platf... Annotations x platf... VideoNot.es: Th... x Opeth - Soldier Of ... Statistics One

www.videonot.es/edit/0B0qVZAFIY8TNjhrEHVYVUdFYkk

VideoNot.es camila.canellas@gmail.com

▶ Video (press Ctrl and space to play/pause)

Untitled notes You have unsaved changes

Change video

0:19 First week lecture

0:27 Based on feedback

Share

ne

As you see from the website, we did this course last year.

Feedback and Support

camila 11:13 27/09/2013

VideoNot.es

Feedback - example

The screenshot displays the PolemicTweet interface. At the top left, the logo 'POLEMIC TWEET' is visible. The navigation menu includes 'Accueil', 'Programme', and 'A propos', with language options for '日本語', 'Français', and 'English'. The main content area is divided into three sections:


- Annotations polémiques:** A message stating 'Vous n'êtes pas connecté. Pour participer, veuillez vous identifier en cliquant ici.' with a set of feedback buttons (??, ==, --, ++, Envoyer).
- Rechercher:** A search bar with a magnifying glass icon and a '1 min.' filter.
- Tweet Stream:** A list of tweets with a vertical timeline on the right. The tweets include:
 - @cybunk: #enmi ne pas épuisé la confiance en écrivant tout -- contrat assurance catastrophe transparence ...
 - @vincentpuig: #enmi André Orlean : l'empire de la valeur. La probabilité statistique se mue en certitude mais pas
 - @nicolasauret: @cybunk RT @GayaneAdourian: J'écoute, je twitte, je storifie, je discute... Manque plus que les
 - @NicolasLoubet (Nicolas LOUBET) #ENMI RT @PolemicTweet Positionner vos tweet avec -- pour désaccord, ++ accord, == référence, ?? questions #Live bit.ly/rU79fl
 - @GayaneAdourian: J'écoute, je twitte, je storifie, je discute... Manque plus que les

On the right side, there are two tabs: 'VIDÉO' and 'SEMANTIC BOARD'. The 'SEMANTIC BOARD' is active, showing a network graph with nodes representing users. The central node is '@nicolasloubet', with other nodes including '@ylakim', '@audrey_bardon', and '@gayaneadourian'. Below the graph is a vertical slider. At the bottom right, there is a section titled 'Entretiens du nouveau monde industriel 2011' with the date '19 décembre 2011' and the session title 'SESSION 1 - HISTOIRE ET ANTHROPOLOGIE DE LA CONFIANCE'. The description mentions a study on trust perspectives in the context of digital and economic-political crisis.

PolemicTweet

App. of an analysis grid - example

Annotating Academic video - 1.0.0 RC Print Logout



0:38

Annotate on My new Track

Layout Collapse

Write a free text annotation. Use »shift + return« keys to create a new line.

Pause video during writing Insert

Mine Edit mode

Landsc	Weather	Part of
Mt Mountain	Sny Sunny	Mrg Morning
Cty City	Cdy Cloudy	Mdy Midday
Vlg Village	Rai Rain	Aft After-no
Hls Hill	Snw Snow	Evn Evening

Create a label Create a label Create a label

Timeline

Filter

Track	0	10	20	30	40	50
Default						
My New Track		Si	Mi			
		Hl			Fr	

Reset zoom < + - > + Add track

List

Items visibility Filter Collapse

>	00:00:10	00:00:16	Hls	0		
>	00:00:10		Sny	0		
>	00:00:26		Mt	0		
>	00:00:38		Fresh Salad!	0		

Analysis - example

Projecting American Empire on Film

ASSIGNMENT

Assignment ✕

Class Responses (38)

01 Bullet Memo-Birth of a Nation

by User Name_392, User Name_619

In a focused paragraph of about 100 (but not more than 125) words, illustrated by at least one (but not more than two) clip(s), totaling no more than four minutes, respond to the following question:

How can Barthes's *Mythologies* help us analyze Griffith's *Birth of a Nation* ?

To fulfill this assignment:

1. Watch *Birth of a Nation* and make clips for your analysis (you may make more selections (clips) than you will use in this bullet memo).
2. From the Home page, scroll down to the assignment title "01 Bullet Memo-Birth of a Nation" and click the green "Respond to Assignment" button; or alternatively, return to this assignment window and click the "Respond" button, and write out your answer, incorporating your already-made selections (clips) using the "Add selection to composition" arrow icon in the top right corner of the selection.
3. Make sure you save your response on the "Published to Whole Class" level to submit it.

VIEW/INSERTED SELECTIONS

COMPOSITION

Published to Class ✕

Create Instructor Feedback


Silent Stereotypes and Barthesian Myths

by User Name_2972

In *Birth of a Nation*, a great deal of the film's spin and message derives from the characters' mythological body language. However, the film seeks to create myths of black men and women in order to create contrasts and further its ultimate message. In the **first extended portrayal of blacks** in the film, their actions and appearances are exaggerated; most of them walk with a hunchback or limp, and clap hyperbolically. This is a method meant to make physical and visual the myth of Black uncivility, a logic repeatedly employed throughout *Birth*. Women are also presented as exaggerated figures — here **Hydia Brown** Stoneman's housekeeper reacts to Stoneman's edict of equality in a sexual manner, suggesting a Barthesian myth of black female promiscuity and manipulation.

cameron meeting slaves
from *Birth of a Nation*

00:14:12/03:31:09



VIEW/INSERTED SELECTIONS

Reflexivity/Feedback - example

The screenshot displays the VISU interface for a session titled "Les loisirs", recorded on 20-10-2011 at 15:14. The interface includes a top navigation bar with buttons for "Accueil", "Utilisateurs", "Séances", "Salon synchrone", and "Bilans". The language is set to "Français", and the user is identified as "S. Serguei".

The main area shows three video feeds: V. Caroline (top left), S. Serguei (top right), and A. Belin (bottom right). Below the video feeds is a timeline with a play button and a current time of 00:05:17. A red circle highlights the 00:05:17 mark on the timeline, which is linked to a comment box containing the text "vous voyer / voier".

The bottom left sidebar contains a "Comments" section and a "Documents" section with options for "Marqueurs", "Messages", "Images", and "Videos". The bottom right section is titled "Liste des bilans".

Visu

Course enrichment - example

The screenshot displays the LeCto LMS interface with several panels:

- Top Panel:** Navigation buttons for "Restart Lecture", "Create PPT file", "Create index", "Write notes", "Write links", "Write web", "Write quiz", and "Write FAQ".
- Left Panel (Radna površina):** A slide titled "Važne teorije i tvrdnje - sažetak" (Important theories and statements - summary) with a list of bullet points. A circled number "1" is next to the first point.
- Middle Panel (Predavač):** A video player showing a lecturer at a whiteboard. A circled number "2" is in the top left corner of the video frame.
- Right Panel (Predavačeve bilješke):** Text describing the lecture method. A circled number "3" is at the bottom left of the text.
- Bottom Panel (Web sažetak):** A section titled "Meaningful learning" with a list of four points. A circled number "4" is at the top left of the text.
- Bottom Left Panel (Kvizi):** A quiz section with two questions. A circled number "5" is next to the first question.
- Bottom Left Panel (Navigacija po slajdovima):** A navigation menu with a circled number "6" next to the "Važne teorije i tvrdnje - sažetak" item.
- Bottom Left Panel (Lisnari):** A list of documents with a circled number "7" next to "Zaključni teoreti učena".

INQUE

Annotation challenges

- Goals
 - Ensure interoperability
 - Ensure durability
- Support
 - Anchoring now normalized (MediaFragment)
 - From unstructured free-text annotations to semantic annotations

Semi-automatic annotation challenges

- Many efforts to do automatic generation (Translectures, linkedTV) but not perfect yet
- Provide tools that **combine** automatic algorithms and correction interfaces
- Being able to cope with hundreds of annotations

NAVIGATION (AND INDEXING)



Cdlh 2014

Linking the segments

- As videos are becoming increasingly basic objects, how do we navigate inside a video?
- And from video to video?



MediaMixer:

Community set-up and networking for the reMIXing of online MEDIA fragments

The main rationale of MediaMixer is to set up and sustain a community of video producers, hosters, and redistributors who will be supported in the adoption of semantic multimedia technology in their systems and workflows to build a European market for media fragment re-purposing and re-selling.

55

Multimodality

- In a first approach, having to deal with material coming from different sources is a (technical) problem
- But it is also an asset!
 - Pdfs give us a lot of information towards language modelling
 - Slides allow us to envisage segmentation

COCo project (in Nantes)

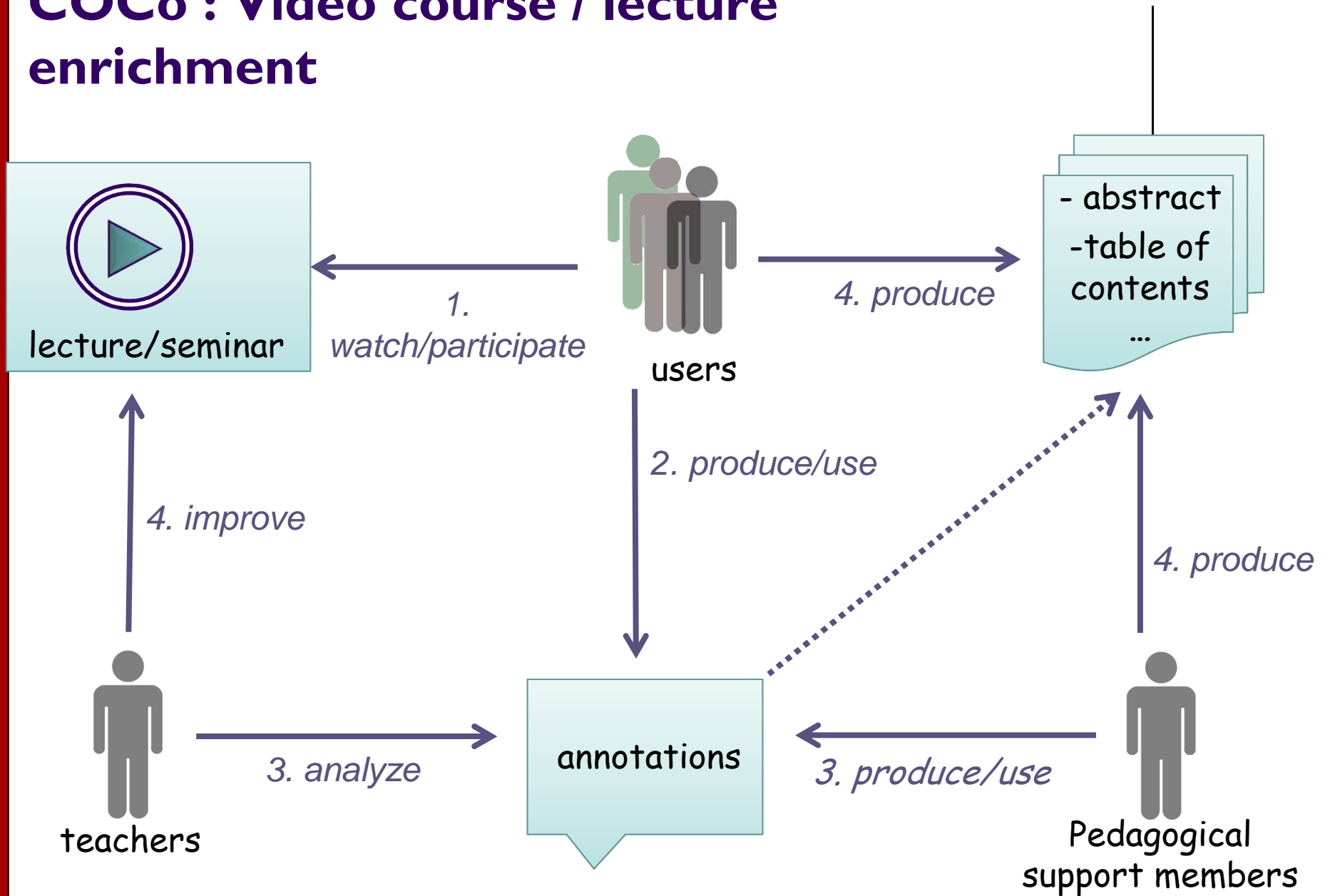
CominOpenCourseware



Cdlh 2014



COCo : Video course / lecture enrichment



Automatic alignment

single named entities (e.g., "harry potter trailer") and less than 1% of the queries contain two or more named entities (e.g., "hate windsle the reader"). The named entities include people, places, organizations, movies, games, books, music, and etc.

There are strong clues for NEM in click-through data. For example, when searching the trailer of the movie "harry potter", people usually form the query "harry potter trailer", and click the results from movie sites like "movies.yahoo.com". The context "trailer" and the website "movies.yahoo.com" give us strong clues that "harry potter" is the name of a movie. One can also observe that movies usually share similar contexts and/or websites. For queries which do not have contexts (i.e., named entities only) or websites (i.e., no clicked URL), we assume that they have null contexts or null websites.

Click-through data is a useful source for NEM. The scale of click-through data is extremely large and thus the data can bring in collective knowledge of Internet users on named entities. Furthermore, click-through data keeps growing rapidly and thus can provide the most up-to-date information on named entities.

4. PROBLEM FORMALIZATION

We formalize NEM from click-through data as the following problem.

The input to NEM includes a large click-through data set and lists of seed named entities. Each instance in the click-through data is a pair of query and clicked URL, e.g., ("harry potter trailer", "movies.yahoo.com/movies/..."). Each list of named entities corresponds to one class. For example, the list of books may contain "harry potter", "the long tail", etc, while the list of movies may contain "harry potter", "kung fu panda", etc.

We first create a seed data set with the click-through data and named entity lists. Specifically, we scan the click-through data using the given named entities and obtain all the click-through data containing the named entities. We further group the obtained click-through data by named entities. From each click-through instance of a named entity, we extract the context of the named entity from the query (We employ a heuristic rule of taking the rest of the named entity as context. For example "h trailer" from "harry potter trailer", where θ is a place holder), and the website from the clicked URL. Note that the context and the website can be null. The classes to which a named entity belongs are also assigned to the named entity. Note that an entity can be in multiple lists and thus belong to multiple classes. In other words, ambiguities exist in named entity classes. The seed data generation process is depicted in Fig. 1, and Table 1 shows examples of generated seed data.

Named Entity	(Context, Website)	Class
Harry Potter	{ θ book, amazon.com}	Book
	{ θ walkthrough, cinema.go.com}	Movie
	{ θ trailer, imdb.com}	Game
Kung Fu Panda	{ θ best books, wuji}	Book
	{ θ DVD, amazon.com}	Movie
	{ θ trailer, apple.com}	Game

Table 1: Examples of Seed Data

... the seed data, we can, in principle, run a bootstrapping process to mine named entity knowledge. Specifically, we repeat the following two steps: (1) mining new contexts and websites for each class from known named entities of the class, (2) mining new named entities for each class using known contexts and websites of the class. For example, "trailer" and "movies.yahoo.com" can be mined for the movie class from the seed data. The patterns can then be used to mine new movie names.

The biggest challenge here is how to deal with the class ambiguities of named entities. For example, "harry potter" are both book and movie, and thus the contexts and websites of both book and movie are associated with it in the click-through data. Taking a deterministic approach would be difficult to deal with the ambiguity. However, we define and utilize a topic model.

5. MODEL AND ALGORITHM

5.1 Topic Model

Topic model naturally gives rise to words in documents. Each document is associated with a set of topics, and each topic is associated with a set of words. Words in documents are generated from topics.

Here we define a topic model to generate click-through data. In the topic model, topics represent the classes of named entities, virtual words represent context-website pairs, and each virtual document is associated with a named entity. Note that context-website pairs are actually co-occurring two types of words. Words in a document (context website) are associated with a named entity is assumed to be probabilistically determined by the topics of the document (classes of the named entity). The difference between topic model of click-through data and topic model of documents is depicted in Fig. 2.

The generative process of the topic model corresponds to the process of search. The searcher first decides a named entity with specific class to search for, then formulates the query (i.e., picks up context), and clicks relevant result (i.e., selects website). For example, if a searcher looks for "harry potter" movie, he would form queries with movie contexts, such as "trailer" and "dvd", and prefers the search results from movie websites, like "imdb.com". If he is interested in "harry potter" book, he would include book contexts, such as "summary" and "notes", and tends to click the results from book websites, like "sparknotes.com". Different named entities have different distributions over classes, for example, entity "harry potter" has a high probability on "movie", while "hate" has a very high probability on "game".

Latent Dirichlet Allocation

- Deal with ambiguity in classes of named entities
 - Classes of named entities are ambiguous.
 - Harry Potter, Book, Movie and Game
 - Topic models (LDA)
 - Deal with ambiguity in classes of named entities

... this figure <sil> on shows the intuition behind this choice <sil> so in the figure <sil> classes in <sil> at the and represents <sil> true topics <sil> and whether the user's consider her report <sil> as <sil> the movie title <sil> they would like to <sils> the of are likely to form <sil> formal such as part of all the true and <sil> I recall the DVD <silence>and becomes <sil> sites such as I <sil> outcome <sil> of always thought yeah for course <sil> on the other hand <sil> when the user conceded Harry Potter as <sil> given <sil> of the probably <sil> a form of a storey <sil> such as...

TRACKS (LEARNING ANALYTICS)



Cdlh 2014

Daphne Koller

- You can turn the study of learning from the **hypothesis driven** mode to the **data driven** mode... a transformation that has for example revolutionized biology
- http://www.ted.com/talks/daphne_koller_what_we_re_learning_from_online_education

A very small conclusion

- There is a lot of research aiming to improve the technologies linked with online education
- Machine learning is a key (but not unique) technology
- More at OCWC 2014 Global!

Thank you.

WEBSITES:

<http://www.k4all.org/>

<http://videolectures.net/>

<http://www.translectures.eu/>

Knowledge 4All Foundation Ltd



Sugata Mitra

A teacher who can be replaced by a computer, should be.

<http://sfltdu.blogspot.fr/2012/11/a-teacher-who-can-be-replaced-by.html>