NLPKD Day

Wednesday May 31, 2017 (08:50-12:30)

A008

08:50: **Nyoman Juniarta**, *Orpailleur* — "A dynamic and multidimensional recommender for managing cross-cultural heritage in Europe"

09:00: **Thien Hoa Le**, *Synalp* — "Does Convolutional Network need to be Deep for Text Classification?"

09:10: **Kevin Dalleau**, *Orpailleur* — "Graph aggregation for classification purposes : application to disease nosography"

09:20: **Younes Abid**, *Orpailleur* — "Assessing privacy leak on social networks"

09:40: **Quentin Brabant**, *Orpailleur* — "From RDF Data to Multi-Level Pattern Structures"

10:00: **Justine Reynaud**, *Orpailleur* — "Classifying the Content of the Web of Data Based on Formal Concept Analysis and Pattern Structures"

10:20: **Emilie Colin**, *Synalp* — "Natural language processing and automatic generation of grammar exercises"

10:30-11:00: **Coffee break**

11:00: **Anastasiia Tsukanova**, *Multispeech* — "Articulatory speech synthesis from static MRI captures"

11:20: **Mathieu Fontaine**, *Multispeech* — "Parameterized Wiener filter with alpha-stable processes"

11:30: **Pierre Mercuriali**, *Orpailleur* — "Median based computing and median structures for classification"

11:40: **Pierre Monnin**, *Orpailleur* — "Extracting and comparing knowledge in pharmacogenomics"

11:50: **Sara Dahmani**, *Multispeech* — "Expressive Text Driven Audio-Visual Speech Synthesis"

12:00: **Zeinab Bakhtiarinoodeh**, *Cello* — "Reasoning with incomplete and inconsistent information"

12:20: **Théo Biasutto-Lervat**, *Multispeech* — "Multimodal coarticulation modelling"

Nyoman Juniarta

A DYNAMIC AND MULTIDIMENSIONAL RECOMMENDER FOR MANAGING CROSS-CULTURAL HERITAGE IN EUROPE

Orpailleur

08:50-09:00

CrossCult is a multinational project about European cultural heritage. One of their objectives is to improve the quality of museum visitors around Europe. This can be achieved by studying the database of visitor trajectories in a museum. Each trajectory is treated as a sequence of museum's artifacts. Therefore it can be considered as a sequential pattern mining problem, which try to find frequent sequences among a set of sequences. In the first part of our work, we attempt to simplify this problem using sampling method.

In a related and simpler problem, namely frequent itemset mining, there exists sampling procedures for selecting interesting patterns, so the complete enumeration of all itemsets can be avoided. Instead of randomized sampling, these procedures have a control over the distribution of the samples, according to the size of each itemset in a dataset. We try to extend these procedures into sequential pattern mining, so the time and space complexity will be hopefully decreased, without significantly reducing the quality of discovered patterns.

Thien Hoa Le

DOES CONVOLUTIONAL NETWORK NEED TO BE DEEP FOR TEXT CLASSIFICATION?

Synalp

09:00-09:10

Convolutional Network now becomes ubiquitous on many Image Classification tasks because it can retrieve the state-of-the-art performance when it goes very deeply. The same effect has been observed in Speech Recognition but is it always the case for Text Classification ? There are a lot of results against this suspect. In this presentation, we will provide the first empirical demonstration to support this fact. The direct consequence will result in subsequent study of the deep network structure for text and its application in many NLP tasks.

Kevin Dalleau

GRAPH AGGREGATION FOR CLASSIFICATION PURPOSES : APPLICATION TO DISEASE NOSOGRAPHY

Orpailleur

09:10-09:20

Graphs have a great expressive power enabling the representation of complex data, as well as their relations and the application of powerful methods. In the setting of the *Fight Heart Failure* project, the strong heterogeneicity of the biomedical data we wish to analyse and the myriad of relations between them brings us to consider the labeled property graph model as a representation. We aim at defining aggregation operators on these graphs, with three goals in mind :

1. Define similarity measures between two graphs, enabling the comparison of patients using their respective graphs,
2. Using these newly defined similarity measures to leverage powerful classification methods, supervised and unsupervised,
3. Enable insightful data visualisation from these labeled property graphs.

The application of these aggregation methods on patient data can lead to a better classification of the diseases they suffer from, *i.e*, heart failure in this very specific case.

Younes Abid

ASSESSING PRIVACY LEAK ON SOCIAL NETWORKS

Orpailleur

09:20-09:40

A well known fact is that social networks leak information, that may be sensitive, about users. However, performing accurate real world online privacy attacks in a reasonable time frame remains a challenging task. We address the problem of rapidly disclosing many friendship links using only legitimate queries. Our study sheds new light on the intrinsic relation between communities (usually represented as groups) and friendships between individuals. Our proposed algorithm is able to perform friendship and mutual-friend attacks along a strategy that minimizes the number of queries. The results of attacks performed on active Facebook profiles show that 5 different friendship links are disclosed in average for each single legitimate query in the best case.

In order to demonstrate privacy threats in social networks we show how to infer user preferences by random walks in a multiple graph representing simultaneously attributes and relationships links. For the approach to scale in a first phase we reduce the space of attribute values by partition in balanced homogeneous clusters. Following the Deepwalk approach, the random walks are considered as sentences. Hence unsupervised learning techniques from natural languages processing can be employed in second phase to deduce semantic similarities of some attributes. We conduct experiments on real datasets to evaluate our approach.

Quentin Brabant

FROM RDF DATA TO MULTI-LEVEL PATTERN STRUCTURES

Orpailleur

09:40-10:00

The recent development of the web of data made available an increasing amount of ontological knowledge in various fields of interest. This knowledge is mainly set out in the RDF and RDFS specifications, and thus takes the form of subject-predicate-object triples. Each triple expresses that the subject is related to the object in a sense that is specified by the predicate value.

We aimed to apply Formal Concept Analysis (FCA) and pattern structures to this kind of data. FCA is a framework for knowledge discovery in which every object is associated to a description. It allows to compare and classify objects with respect to their descriptions, but also to highlight dependencies between different parts of the descriptions. Pattern structures can be seen as an extension of FCA, to deal with complex descriptions in an easy way. FCA and pattern structures rely on the notion of similarity. Roughly speaking, the similarity is an operation that takes two descriptions and generates a new one that contains common characteristics of both.

Here, the objects we consider are certain entities from an RDF base. The main problem lies in the definition of the descriptions of these objects, and of the similarity operation; how can we represent relevant information about each entity from the RDF base, and then, how can we compute the similarity between two descriptions in a meaningful way?

Justine Reynaud

CLASSIFYING THE CONTENT OF THE WEB OF DATA BASED ON FORMAL CONCEPT ANALYSIS AND PATTERN STRUCTURES

Orpailleur

10:00-10:20

The Web of Data has become a very huge space of experimentation especially regarding knowledge discovery and knowledge engineering due to its rich and diverse nature. Furthermore, the Web of Data is based on RDF triples of the form <subject, predicate, object> where each element in the triple denotes a resource (accessible through a URI). Moreover, the elements in a triple can be organized within partial orderings using the predefined vocabularies such as RDF Schema (RDFS), i.e. a subclass relation (\texttt{rdfs:subClassOf}) and a subproperty relation (\texttt{rdfs:subPropertyOf}, where a predicate in an RDF triple is also called a property).

This presentation will focus on a framework based on Formal Concept Analysis and the Pattern Structures (PS) for classifying sets of RDF triples. Formal Concept Analysis (FCA) is a mathematical framework used for classification and knowledge discovery. As a learning process, FCA allows to build an ordered set of concepts where objects are classified w.r.t. the attributes that they share. Pattern structures are a generalization of FCA for dealing with complex data. First, a pattern structure extending previous approaches will be defined. Then, experimental results from Dbpedia will be presented.

Emilie Colin

NATURAL LANGUAGE PROCESSING AND AUTOMATIC GENERATION OF GRAMMAR EXERCISES

Synalp

10:20-10:30

Aims

My thesis aims to explore deep learning methods to generate correct and fluent sentenses. It is part of the METAL collaborative project on e-learning. My current research goal is to generate under constraints. For example, given an input consisting of some keywords describing an event (*cat* (subject) *eat* (relation) *mouse* (object)) and of a syntactic constraint (*passive needed*), the aim is to automatically generate a sentence such as *Mice are eaten by the cats*. Given a more complex input (several subject/relation/object triples) allows for syntactic constraints such as object— or subject— relative clause… For instance, given the input mouse eat cheese/cat eat mouse and a subject relative clause constraint, the generated sentence might be : T*he cat eat the mouse that eat cheese*.

The motivation for generating under constraints is that it can be used to generate sentences that are lexically and syntactically appropriate to automatically generate grammar exercises.

Method

Recently, sequence to sequence models have met success for translation and text generation. The sequence to sequence models were introduced by Cho *et al.*[2] in 2014, and successfully used for obtaining translated sentences by Sutskever *et al*.[3], also in 2014. Wen *et al*.'s work [4] uses an LSTM[1]-based system to generate under semantic constraints, in a very restricted dialog domain. My goal is to explore how such models can be adapted to generate under constraints.

As a first step towards that goal, I am currently working on building a training corpus using the WebNLG dataset (semantic web data associated with lexicalizations) and enriching the input data with syntactic labels indicating the syntactic constructs occurring in the corresponding sentence.

Références

[1] Fayol, Michel and Jaffré, Jean-Pierre, *L'orthographe.* Presses universitaires de France, 2014.

[2] Kyunghyun Cho and Bart van Merrienboer and Çaglar Gülçehre and Fethi Bougares and Holger Schwenk and Yoshua Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, CoRR, 2014.

[3] Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V, Sequence to sequence learning with neural networks, Advances in neural information processing systems, 2014

[4] Wen, Tsung-Hsien and Gasic, Milica and Mrksic, Nikola and Su, Pei-Hao and Vandyke, David and Young, Steve, Semantically conditioned lstm-based natural language generation for spoken dialogue systems, arXiv, 2015.

Anastasiia Tsukanova

ARTICULATORY SPEECH SYNTHESIS FROM STATIC MRI CAPTURES

Multispeech

11:00-11:20

In this talk I will present my work that has been carried out in the domain of articulatory speech synthesis — synthesizing speech through simultaneous control over the

[1] LSTM , Long Short-term Memory

articulators (the jaw, the tongue, the lips, the velum, the larynx and the epiglottis) and the source — based on static MRI data (97 annotated images capturing the articulation of French vowels and blocked consonant-vowel syllables). This rule-based control has to take into account coarticulation and be flexible enough to be able to vary strategies for speech production. The results of this synthesis are evaluated visually, acoustically and perceptually, and the problems encountered are broken down by their origin: the dataset, its modelling, the algorithm for managing the vocal tract shapes, their translation to the area functions, and the acoustic simulation.
Then I will describe the recently acquired dynamic data (rtMRI) and talk about our plans regarding it.

Mathieu Fontaine

PARAMETERIZED WIENER FILTER WITH ALPHA-STABLE PROCESSES

Multispeech

11:20-11:30

We introduces a new method for single-channel denoising that sheds new light on classical early developments on this topic that occurred in the 70's and 80's with Wiener filtering and spectral subtraction. Operating both in the short-time Fourier transform domain, these methods consist in estimating the power spectral density (PSD) of the noise without speech. Then, the clean speech signal is obtained by manipulating the corrupted time-frequency bins thanks to these noise PSD estimates. Theoretically grounded when using power spectra, these methods were subsequently generalized to magnitude spectra, or shown to yield better performance by weighting the PSDs in the so-called parameterized Wiener filter. Both these strategies were long considered ad-hoc. To the best of our knowledge, while we recently proposed an interpretation of magnitude processing, there is still no theoretical result that would justify the better performance of parameterized Wiener filters. Here, we show how the alpha-stable probabilistic model for waveforms naturally leads to these weighted filters and we provide a grounded and fast algorithm to enhance corrupted audio that compares favorably with classical denoising methods.

Pierre Mercuriali

MEDIAN BASED COMPUTING AND MEDIAN STRUCTURES FOR CLASSIFICATION

Orpailleur

11:30-11:40

Together with Miguel Couceiro and Romain Péchoux, we have been working on models of computation that involve the median operator and study their efficiency compared to other well-known models such as the Disjunctive, Conjunctive, and Zhegalkin (Reed-Muller) normal forms, with applications in decision theory and agregation.

In particular, in the framework of Boolean functions, we have compared various normal form systems (NFSs) in terms of the asymptotic efficiency of the representations they produce. We identified some properties such as associativity, linearity, quasi-linearity and symmetry, that allow the corresponding NFSs to be compared. A noteworthy result we have obtained that is tied to associativity is that systems generated by a single connective (for instance, the Sheffer stroke $\neg(A \wedge B)$) are always polynomially as efficient as any system generated by two or more connectives: using more connectives does not yield more efficient representations. From this result and others we have shown that the median normal form, that has the ternary median as its only non trivial connective, is polynomially as efficient as any other NFS.

Another area of study is a median-based calculus to represent polynomial functions over distributive lattices efficiently. We have given ways to simplify the resulting median formulas and investigated related complexity issues: for instance, the problem of deciding whether a median formula is the smallest, according to a certain measure, is in Σ^P_2.

Pierre Monnin

EXTRACTING AND COMPARING KNOWLEDGE IN PHARMACOGENOMICS

Orpailleur

11:40-11:50

Pharmacogenomics studies the influence of the genome in drug response. Particularly, it aims at identifying relationships between genomic variations and variability in drug response, including adverse drug effects. Pharmacogenomics is one of the main components of personalized medicine which consists in tailoring drugs and doses to a patient’s genome in order to maximize drug expected effects and minimize risks of adverse reactions. A huge number of pharmacogenomic relationships are available in literature and specialized knowledge bases. Some of them have been significantly studied and confirmed while others have only been observed on reduced cohorts and remain to be further investigated. On the other hand, nowadays, lots of health care data are digitally available thanks to the use of Electronic Health Records (EHRs). They contain information about diseases, laboratory tests, medical procedures and prescriptions that a patient has experienced. In this talk, I will present the PractiKPharma project which aims at mining EHRs to confirm or temper pharmacogenomic relationships and automatically explaining pharmacogenomic mechanisms. I will particularly focus on the challenges of my PhD thesis: extracting biomedical knowledge from EHRs and comparing it to referential knowledge.

Sara Dahmani

EXPRESSIVE TEXT DRIVEN AUDIO-VISUAL SPEECH SYNTHESIS

Multispeech

11:50-12:00

Talking avatar has been widely used in many human-computer interaction in the last decades. Virtual hosts and embodied intelligent agents must be natural to make the interaction experience comfortable for humans.

Besides the multimodal (AudioVisual) aspect of the speech, emotional facial expressions can further enhance interaction through non verbal communication. In this talk, I will explain the main difficulties that we face when integrating emotions in speech and how we plan to address them.

Zeinab Bakhtiarinoodeh

REASONING WITH INCOMPLETE AND INCONSISTENT INFORMATION

Cello

12:00-12:20

In the past decades, reasoning about knowledge and information change has gained a prominent place in various areas of artificial intelligence and computer science. In these areas, agents have to deal with incomplete and inconsistent information. For example, in distributed systems, agents receive information from multiple sources that may be inconsistent. Moreover, in real-world situations, agents do not have complete information about all aspects of the world and their reasoning power is bounded by thresholds such as time and limited memory. In such circumstances, reasoning about information can be intricacy and demands a careful formal analysis.

Among a number of approaches that has been proposed to model reasoning about information, logic is a prime candidate to do this task.
In this talk, I present a logical framework to formalize reasoning and dynamic aspect of inconsistent and incomplete information.

Théo Biasutto-Lervat

MULTIMODAL COARTICULATION MODELLING

Multispeech

12:20-12:30

Speech is the natural communication medium between human beings.
Thus, an animated talking head with realistic speech movements should be an effective human-machine interface. Moreover, it's well established that viewing the speaker's face tends to increase speech intelligibility. Hence such technology could have several applications: communication aids for hard-of-hearing people, embodied conversational agents for noisy places, and more. However, proper computation of speech gestures should take account of coarticulation phenomenons, i.e. the influence of a sound and its articulation on another phoneme's articulation. This influence can be both retentive and anticipative, e.g. the lip movement associated with /s/ is different in 'see' and 'sue' because of the anticipation of the protrusion needed by /u/.
Motivated by the recent success of deep learning techniques in many language-related tasks, we try to model these coarticulation effects with a neural network approach. As coarticulation is highly temporal-dependent, recurrent neural networks seem to be a tool of choice by their capacities to deal with time series. In a supervised fashion, we learn to generate the articulatory trajectories given the phoneme sequence and some acoustic features.