Xerox Research Centre Europe
1993 - 2013
20 Years of Innovation in Europe
This book was created for the 20th anniversary celebration of Xerox Research Centre Europe. We hope you enjoy it.
1993 - 2013

Celebrating two decades of turning scientific vision into business reality.
Foreword

Twenty years ago the world was a very different place. Email was reserved for the privileged few, mobile devices were heavy, clunky and expensive while computers were primarily used for word processing. We were about to undergo a technological revolution and, although Xerox could not predict exactly what shape it would take, those at the helm were sure of two things. First, that software would differentiate hardware and that innovation was where that difference would come from.

I will never forget the excitement of being part of a team given the task of creating a research centre, in computing, in France to build such a future. We began by establishing a network with the scientific community throughout Europe and in the Rhône-Alpes region, our prime choice of location because of its strengths in academia and industry. In partnership we created multilingual information management tools and distributed computing systems for the growing EU community and market. We developed applications that ran on networks, on computers and mobile and printing devices. To respond to the need and desire to share those applications with our customers we created the first Xerox Technology Showroom and demonstrated the Office of the Future in 1995.

The first commercialisation of our research was in 1996 with technology for the professional translation market. When Xerox expanded into the broader networked document services space we were ready to provide technology in Content and Document Management bridging the paper, digital and mobile worlds. In parallel with the development of new software architectures and services we expanded our suite of imaging and text tools which became known throughout and beyond the company walls as the Smarter Document ManagementSM technologies. Today these are used to process 30 million documents every month in our customer scanning sites and to categorize millions of documents in the preparation of litigation cases in court.

Three years ago Xerox acquired a leading business process services provider called Affiliated Computer Services. This acquisition dramatically changed our business and was well in-line with the services research agenda that XRCE had been assiduously pursuing. I was fortunate to be given the privilege to realign our services research around the world to the new requirements. Annual revenue from services is now greater than that from technology and it covers numerous areas we had never previously explored.
Our research at the centre in data, text and imaging is now being applied to innovate not only in document devices but in New Services Verticals in Transportation, Healthcare and Customer Care.

Today, the excitement that reigned at the beginning of the centre is just as strong and the impact even greater. I am confident you will feel both in the pages that follow and in the exchanges you have with us.

Sincerely,

Monica Beltrametti
Xerox Chief Services Research Officer
Vice President and Director Xerox Research Centre Europe
# Table of Contents

**Foreword** 6

**Celebration October 4\textsuperscript{th}, 2013** 11
- Overview 11
- Emerging Trends Presentations 12
- Technology Demonstrations 14
- Distinguished Seminar Series 20

**Anniversary Articles** 23

**20 Years of Innovation Timeline** 64

**Xerox Research Centre Europe** 66

**Activity Overview** 70
- Natural Language Processing and Information Retrieval 70
- Computer Vision 76
- Collaborative and Interactive Systems 82
- Data Analysis and Process Modelling 88

**Our People** 94

**Technology Showroom** 96

**Open and Collaborative Innovation** 98
Xerox Research Centre Europe

20 Years of Innovation
# Agenda  
Friday, 4\textsuperscript{th} October, 2013

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30</td>
<td>Welcome</td>
</tr>
<tr>
<td></td>
<td>Monica Beltrametti, \textit{Xerox Chief Services Research Officer, Vice President and Director Xerox Research Centre Europe}</td>
</tr>
</tbody>
</table>

**Expanding European Markets**

Jacques Guers, \textit{President, Xerox Europe}

**Industrial Research in France**

Geneviève Fioraso, \textit{Minister for Research and Higher Education, French Government. Representative of the 1\textsuperscript{st} district of Isère}

<table>
<thead>
<tr>
<th>10:00</th>
<th>Panel Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How should the focus of IT research evolve to have the greatest impact on people and organizations?</td>
</tr>
<tr>
<td></td>
<td>Moderator: Katie Ledger</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11:00</th>
<th>Emerging Trends Presentations (Pages 12 -13)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Four 15 minute presentations that will take you into the future.</td>
</tr>
</tbody>
</table>

**Technology Demo Fair** (Pages 14 -19)

Researchers demonstrate how different technologies are being applied to real business cases.

| 12:30 | Cocktail and Lunch |
Emerging Trends Presentations

11:00 - 12:30 Mont Blanc room (Quartz building)

A series of 15 minute future presentations in a dedicated theatre.

11:10 - 11:25
The future of the workplace: through an ethnographer’s eyes.
Jacki O’Neill

What will the workplace of the future look like? Will it even exist as a physical space? Ever since networked computers first became widespread the idea of the virtual organisation has been touted, yet most of us still work in conventional workplaces. Recent technological advances have given a second wind to the idea with some even questioning whether organisations will exist in their current form or whether workers will be free agents coming together to complete specific tasks as and when needed. Discover the answer through the eyes of an ethnographer who spends their time...studying people in workplaces.

11:30 - 11:45
How social mass communication and language technology will change the way we run society.
James Henderson

The advent of social media has allowed everyone to broadcast their opinion, but this rarely leads to the kind of common, shared understanding provided by traditional mass media. On-going advances in language technology that automatically summarise large sets of opinions will enable this shared understanding of the public. The result will change the way people communicate, and with it the nature of society.
Self-driving cars: on the eve of a technological revolution

Boris Chidlovskii

The introduction of “self-driving” cars will dramatically reshape not only the landscape of the automotive industry but also the role vehicles play in our lives and the future design of our roads and cities. Existing solutions have their limitations but will be overcome by the eventual convergence of sensor-based and communication-based vehicle technologies. The service opportunities that will arise from their widespread adoption are endless, in particular those based on the huge volume of data produced and the need to marshal heterogeneous data sources.

From culture to culture: the new frontier in a volatile digital world.

Jean-Pierre Chanod

The most visible aspect of today’s digital world is the explosion of content and data. Yet as long as this information remains siloed and inconsistent there is only so much one can do with it. This disconnected dispersion of knowledge will continue to accelerate with the increase of new data and information resulting in a greater fragmentation of society. To overcome this schism we need systems that make it possible for people to interact and collectively elaborate knowledge that goes beyond each of their particular realms to realise unprecedented discoveries. We call this a ‘computer assisted Renaissance’. For such information systems to appear we must first break free from the volatility and obsolescence inherent in digital media.
Technology Demonstrations: Innovation in Action

11:00 - 12:30 Everest room and Technology Showroom (Château building)

XRCE innovators will show interactive demos how innovation is being used in transportation, customer care, document services and other areas.

Pod 1
Room Everest

Making it easier to park with dynamic pricing. Onno Zoeter

Looking for somewhere to park on city centre streets is one of the main causes of urban congestion and pollution. Xerox has developed an analytics-based engine that dynamically adjusts parking rates based on driver demand for spaces and availability. The parking engine has been integrated as a module into Xerox’s new Merge™ parking management system, which is a single portal for managing a city’s metres, pay-by-mobile phone, sensors, enforcement and collections. It is currently deployed in Los Angeles.

Pod 2
Room Everest

Using analytics to make public transportation networks more effective. Frédéric Roulland

As a passenger, every time you swipe your credit card or ticket you let the operator know when and where you started your journey yet, instead of making use of such valuable information, it often lies around unused. Xerox expertise in analysing printer data is being used on ticketing data to give city transportation agencies a better understanding of how travellers actually use public transit, enabling government organizations to better respond to customer needs. The system is being tested in several cities around the world.
New generation license plate matching.
José-Antonio Rodríguez-Serrano

License Plate Matching is needed in multiple transportation applications: Parking flow management, Entry/Exit monitoring, Point-to-Point enforcement (such as Average Speed Enforcement), and as an additional identity verification level in Electronic Tolling. Existing license plate recognition systems present difficult challenges exist when it comes to increasing the recognition rates of license plates or protecting the identity of the driver. The vehicle Re-identification Technologies developed at Xerox Research Centre Europe allow improved matching of vehicle identities across points of a network, while respecting the privacy of drivers.

Retrieving the right image from large-scale image databases. Florent Perronnin

With the availability of billions of digital media assets, from photos to videos or marketing collaterals, users demand better and more powerful tools to help them store, organize or manipulate such data. The image retrieval system demonstrates perfectly the power of Xerox image representation and compression algorithms. Capable of searching five million images in less than a second on a standard desktop computer (no need for cloud), users can search large collections of high quality photos, digital assets, graphic design elements or scanned documents. It can be used to help organize, visualize and query all types of image-based repositories, from large photo stock databases, enterprise scale digital document repositories or a personal photo collection on your desktop.
Pod 5

Room Everest

**Xerox Agent Performance Indicator.**
François Ragnet

Designed to boost call centre agent motivation, this technology provides every employee with a mini dashboard providing a personal real time performance status. Using gamification techniques, call centre managers can start games to stimulate agents. Overall the tool provides better agent satisfaction which has a direct positive impact on client experience.

Pod 6

Room Everest

Towards virtual call centres: automatically analysing customer requests. Denys Proux

Providing the best client experience is the goal of Xerox call centre agents. This means finding the best answer to a customer request, in a timely fashion using the right mode of communication. By automatically analysing and indexing client calls helps identify if the current customer request has already occurred and how it was solved. This shortens the time necessary to provide the solution and also improves the quality of response and the overall customer experience.
Pod 7

Room Everest

Machine translation improves customer response times. Jean-Luc Meunier

You don’t even have to be global in today’s connected world to receive customer comments and queries in many different languages. When requests come in by email, irrespective of where we are, our agents use automatic machine translation to understand what that query is about. Once they have processed it and are ready to reply, they use our unique multilingual authoring tool to write the response in the customers’ native language. With automatic translation, agent workloads are more easily balanced and responses are faster because dedicated, language-specific agents are not required.

Pod 8

Room Everest

Open Xerox. Nicolas Guérin

Open Xerox is a web portal that hosts technology prototypes from the Xerox R&D labs, making them accessible to the external user community well before the launch of a product offering. Open Xerox is also a gateway to information on other interesting Xerox lab projects, or to (open source) software that can be downloaded and executed on a user desktop, with instructions user manuals and starter kits.
Pod 9
Technology Showroom

Virtual Help Desk. Yves Hoppenot.

This award-winning technology offers virtual help via 3D diagnostics and instant live support directly from a printer. It works by pressing the ‘Help’ button on the screen of a networked. A recent addition to this solution makes it possible to connect to the printer from a mobile device such as a tablet so the user can troubleshoot without calling an agent. The solution results in increased up-time and customer satisfaction.

Pod 10
Technology Showroom

Motivate print users to be greener. Jutta Willamowski

Companies want their employees to print “better”: motivating employees to do so is a real challenge. “Print Awareness Tool” is a self-motivating tool appearing on everyone’s desktop as a flower losing petals if a print behaviour is not environment friendly, typically if one prints simplex instead of duplex, or prints emails in excess. Everyone can assess his own print behaviour, benchmark with peers and is gratified with a « green points » currency when his behaviour improves.
Pod 11

Technology Showroom

Smart Document Review. Caroline Privault

The “smart document” review system allows users to easily work with documents on a large touch-sensitive table top display. The software assists users to quickly group similar documents, and find relevant information. The user can position them side-by-side for comparison, scale them up or down, or pile them in a corner of the table, just as real sheets of papers are placed on a real desk, but with all the capabilities of digital documents. By sorting a few documents as relevant or non-relevant, the user teaches the system what kind of information is of interest. This allows it to then automatically retrieve and suggest new relevant documents dramatically accelerating document processing in areas like litigation, human resources or claim management.
Distinguished Seminar Series

1993 - 2013. In complement to the XRCE weekly technical seminars, this year the centre is hosting several distinguished speakers. These thought leaders will present their current leading edge research and studies to a broad audience. More information and recorded videos of these and all our seminars can be found on our web site at www.xrce.xerox.com.

Upcoming seminar for which you can register via our web site:

4th October 2013, 11:00

The exact title of this talk will be published on the XRCE web site early October.

David M. Blei
Associate Professor in the Computer Science department of Princeton University, U.S.A.

1st October 2013, 17:00

Deep Learning: Machine Learning Via Large-scale Brain Simulations

Andrew Ng
Associate Professor, Stanford University, Stanford, CA, U.S.A.
Ng has also been named to the 2013 “Time 100” list of the most influential people in the world.

Machine learning is a very successful technology, but applying it to a new problem usually means spending a long time hand-designing the input features to feed to the learning algorithm. This is true for applications in vision, audio, and text. To address this, researchers in machine learning have recently developed “deep learning” algorithms, which can automatically learn feature representations from unlabeled data, thus bypassing most of this time-consuming engineering.
JUGAAD: A Frugal and Flexible Approach to Innovation

Navi Radjou

Navi Radjou is an innovation and leadership strategist based in Silicon Valley. He is also a Fellow at Judge Business School, University of Cambridge, and a World Economic Forum (WEF) faculty member. He is a member of WEF’s Global Agenda Council on Design Innovation and a regular columnist on HBR.org.

The traditional Western innovation model, built on costly R&D projects and rigid and time-consuming processes, is no longer sustainable in today’s unpredictable and resource-constrained business environment. Navi Radjou will explain how developing nations (Africa, Brazil, India, China) provide an alternative approach to innovation that is frugal, flexible, and inclusive. This new “street smart” approach, called jugaad innovation, is practiced by thousands of ingenious entrepreneurs in emerging markets who create affordable and sustainable solutions that deliver more value to consumers at less cost while minimizing resource utilization. Western nations have much to learn from emerging economies on how to innovate frugally and flexibly under severe constraints.

Using many real-life case studies of pioneering Western firms such as Siemens, GE, Renault-Nissan, Unilever, Navi will show how you can apply the principles of jugaad innovation within your organization to innovate faster, better, and cheaper.

Modelling Visual Recognition

Jean Ponce

Jean Ponce is Professor and head of the Computer Science Department at the Ecole Normale Superieure (ENS), Paris. He also leads the ENS/INRIA/CNRS project-team WILLOW on computer vision and machine learning.

Modelling visual recognition: This talk addresses the problem of automated visual recognition, that is, having a computer decide whether an instance of some object class, for example, a chair, a person, or a car, is present in some picture, despite shape and colour variations within the class, as well as viewpoint and illumination changes from one photograph to the next. After giving a brief historical perspective on this problem, I will present some recent work aimed at constructing models of both the objects and the recognition process that address different aspects of the problem, including shape variability, finding distinctive parts, and dealing with limited amounts of supervisory information. I will also present some applications to image categorization, object detection, cosegmentation, and video interpretation.

Anniversary Articles

These anniversary articles have been written by research staff from different disciplines with very different backgrounds and roles. They represent a mixture of research projects, challenges and applications. The only common theme is the guideline they were given: ‘Write something you are passionate about!’

Will the next great lost civilization be ours?
Jean-Pierre Chanod, page 22

A tribute to visual words and how they revolutionized computer vision.
Diane Larlus and Florent Perronnin, page 25

Computer science meets journalism.
Matthias Gallé, page 28

Can “made up” languages help computers translate real ones?
Claude Roux, page 31

Do you act like your computer (designer) thinks you ought to?
Jacki O’Neill, page 34

Will translation ever be fully automated?
Marc Dymetman, page 36

How economics and machine learning can tackle transportation congestion.
Eduardo Cardenas, Stéphane Clinchant, Chris Dance and Onno Zoeter, page 39

Do computers need grammar books too?
Âgnes Sándor, page 44

Monks & Markup
Hervé Déjean, page 47

Xerox Living Lab: The customer’s role in our innovation process.
Patrick Mazeau, page 50

How machines can learn what humans interpret: adapting probabilistic topic models to natural language.
Cédric Archambeau, page 54

Medicine leans on natural language technology to advance science and improve patient care.
Denys Proux and Caroline Hagège, page 58

Experience design: the path from research to business.
Yves Hoppenot, Caroline Privault and Antonietta Grasso, page 61
Will the next great lost civilization be ours?

Digital obsolescence is a threat to preserving the records of the 21st century. Information systems of the future will need to be preservation aware by design to ensure the long-term access to and integrity of valuable economic, cultural and intellectual assets.

Clay tablets appeared in Mesopotamia around 2400 B.C. for writing that was meant to last, as opposed to writing on more perishable material such as papyrus. And indeed such tablets did last even beyond the expectations of their authors. Collections of tablets impressed with a stylus in soft clay and then baked in the sun have reached us today, in various languages albeit in different states of conservation. They cover a multiplicity of topics, such as business records, poetry, prayers, hymns, history, divination and science.

Among the most ancient texts recorded on tablets, one may mention the famous Epic of Gilgamesh whose oldest Babylonian fragments date back from the 18th century B.C. or Sumerian philosophical disputations like the Debate between Bird and Fish (2100 B.C.) where the fish tells the bird: “You are shameless: you fill the courtyard with your droppings” while the bird answers “But I am the beautiful and clever Bird! Fine artistry went into my adornment…”

These tablets are striking examples of tangible cultural heritage; where the legacy of physical artefacts inherited from past generations conveys artistic, cultural, religious, documental or aesthetic meaning often produced within a long-gone society.

A tangible heritage is one that can be stored and physically touched. Accessing the meaning behind physical items inherited from the past requires the expertise to recover, translate, compare, interpret and contextualize the text encrypted in the clay tablets or any other ancient artefact such as the infamous Rosetta stone.

The collections of ancient text that still remain to be deciphered are actually huge. The task is so complex and time consuming, that crowdsourcing projects are now emerging, to support experts in this daunting endeavour. For instance, thanks to citizen science projects such as Ancient Lives, non-expert volunteers are invited to help catalogue and transcribe ancient papyri via the Web (http://ancientlives.org). You are welcome to give it a try but be warned, it is not easy!

The tangible traces of ancient civilizations that have reached us today indicate which assets were considered of highest economic or symbolic value at the time, deserving the effort to be engraved on more demanding supports. Their remarkable preservation, even if not originally intended, is a natural side effect of the consideration they received at the time of their creation.

But what about preserving intellectual assets created today? For the most part these exist in digital form, such as documents, emails, images, videos, sound recordings,
computer graphics, websites, sensor data, scientific measurements, medical or legal records. In this digital world, there is no direct equivalent of the tangible objects of the past, such as stone tablets or books. Physically stored data cannot alone be considered as a tangible item in the sense of traditional preservation, since digital content cannot be accessed, and to a large extent, does not exist, without the mediation of a complex computer environment including, beyond physical storage proper, various combinations of hardware and software. Moreover, computer environments change at a rapid pace and are quickly obsolete, making any digital content that relies on a specific environment at a given point in time at risk of soon becoming inaccessible and hence lost. This process is known as digital obsolescence.

Digital obsolescence is a greater threat to the preservation of digital content than the hazards associated with traditional paper documents such as acid, mould and looting combined. Digital obsolescence happens quickly, is pervasive and hard to control. Obsolescence threatens all aspects of the rendering chain, from bits in storage that degrades or for which readers are no longer available, to data formats with outdated documentation or for which the rendering software has disappeared, to software that runs on dead or rare devices and retired operating systems.

Digital preservation is therefore not just about preserving well identified tangible objects, as in the good old days, when maintaining the physical integrity of books, newspapers, manuscripts, pictures, etc. could be achieved with reasonable effort and care, and at a manageable pace. One could always store boxes of documents on long shelves and timely assess, organise and protect the acquired collections. This is no longer possible with digital content. There is little hope to secure durable access without taking specific actions before digital obsolescence comes into play. Soon after digital content has been produced, one must take irrevocable decisions about whether it should be sent to the future and in which form. Otherwise it will be lost forever.

As digital content exponentially grows and digital obsolescence accelerates, preservation will become a major concern for organisations with large data holdings and the need to preserve the critical knowledge contained within. It is anticipated that digital preservation will at some point be seamlessly integrated into the information lifecycle: information systems of the future will be preservation-aware by design. To make this happen, the digital objects of the future will not be treated simply as bit streams associated with adequate hardware and software at their time of creation. They will become part of a rich information ecosystem self-descriptive of all that is essential to know about itself: its purpose, intended behaviour, the context within which it was created, the user experience and more.

To make the descriptions of such ecosystems sustainable, they will be infrastructure-independent, with a strong focus on capturing their temporal evolution and authenticity.

Eventually these descriptions will travel into the future, where yet unknown information systems will need to make sense of them, irrespective of the hardware and software in place at the time when the content was initially created and used. Future generations will then reconstruct not the original digital objects, but new ones. Ones that will convey the essential properties of the originals albeit rendered in a significantly different mode.
This is a major shift in preservation: one no longer preserves tangible physical objects per se, but views or abstract representations of such objects that can be reconstructed in an unpredictable technological future. This shift represents a major challenge for the long term preservation of modern cultural, intellectual and economic assets, the consequences of which are not yet widely recognized. Eventually, digital preservation will become a natural and transparent side effect of the proper governance of information and data.

**Digital Preservation at Xerox Research Centre Europe**

The Xerox Research Centre Europe has been active in the field of digital preservation for over a decade. In the European projects VIKEF, SHAMAN and more recently PERICLES, research addresses various aspects of data and process modelling: metadata creation for collections of digitized books and documents, XML-based document processes, co-design of printed circuit boards, ecosystems of linked IT resources. Beyond digital preservation proper, this research addresses broad challenges that are relevant to the services business at Xerox, such as paper to digital transition, process modelling and data governance.

**About the Author:**

Jean-Pierre Chanod is senior scientist and Area Manager for the Enterprise Architecture group at the Xerox Research Centre Europe. His main research interests include natural language understanding, document processing, data and process modelling.
A Tribute to Visual Words and How they Revolutionized Computer Vision

We love visual content. Think of all the photos we share on social websites. Along with videos, pictures have become a major way of sharing information between individuals. While the power of words is limited by language and cultural barriers, pictures and videos are a universal communication media which transcends such differences.

With the availability of cheap cameras and camera phones, it is not surprising to see visual content overflowing the Internet. As of today, Facebook is estimated to contain more than 100 billion images while on YouTube tens of hours of videos are uploaded every single minute. Yet, this wealth of media and information has little value per se if it cannot be easily accessed.

Tapping into “The Deep Web”

The most natural approach to search for content on the Internet is by issuing text queries on search engines such as Google, Yahoo or Bing. This is because, by far, our preferred mode of interaction with a computer is by using its keyboard. Such interaction makes perfect sense when looking for text. For instance, you will have no problem retrieving the full lyrics of a song from the Internet even if you know only a few words. But there is a huge gap between pictures and words – generally referred to as the semantic gap – which makes the task of matching words and pictures look almost impossible. This is why we tag our pictures and videos and why we give them descriptive names: to be able to retrieve those using text queries. However, hand-tagging visual content is a slow and tedious process and only a fraction of the visual content on the Internet is tagged. The remainder – which is sometimes referred to as the deep web – is inaccessible. We even sometimes have trouble finding pictures in our own personal photo-collection.

What if we could give computers the “gift of sight”? What if we could teach computers to translate our textual queries into “visual queries” so that words can be matched to visual content? But computers see pictures only as a multitude of small colourful dots standing next to each other, i.e. as a matrix of pixels. Why would these colourful dots have any meaning for a machine? The challenge is to link low-level information (the pixels) with high-level concepts like objects and scenes. This is the difficult question that computer vision research tries to address.

Training Computers to “See”

Computer vision is the research field that consists of designing computer programs that help machines understand visual data, for instance, programs that name the objects in a picture, or a video clip. In other words, computer vision scientists teach computers to “see”. They “show” images of different objects to a computer telling them what they are so that it can be trained to recognize these objects – in the same manner one shows images
to a small child, often by pointing at objects and naming them.

While the task of interpreting a scene and its objects is trivial for humans—even for small children—teaching computers to see has proven to be a very arduous task for computer vision researchers. To bridge the semantic gap between the low level pixels and the high-level concepts, it is necessary to introduce intermediate representations. Consequently, the first programs in automatic image understanding have tried to break down the problem of recognizing objects into the problem of recognizing small object pieces. Early methods proposed to decompose objects into geometric components such as cylinders, bricks, wedges or circles. For instance, an ice cream cone is composed of a sphere located above a cone. Although intuitively appealing, such methods obtained moderate success as the problem of recognizing object parts can be as difficult as recognizing the whole object.

As mentioned earlier, pictures are a universal communications medium that is more powerful but more complex than text. As a document is a series of words and words are well defined entities, we can count how often each word appears in a document and represent this document as the number of occurrences of each word. This simple representation—known as the bag-of-words—is very powerful and is at the heart of every modern text search engine. It allows us to classify documents using the presence (or absence) of certain words, as this is a strong indicator of the topic of a document. For instance, words such as “score”, “ball” or “league” are strongly indicative of a sports theme.

**Xerox and the Visual Vocabulary**

What if we could define entities made up of pixels such as “visual words”, so that similar representations could be used to describe images? This is exactly what Xerox scientists managed to achieve and, in the process, they revolutionized the field of computer vision. By analogizing visual content with textual content they introduced the concept of a “visual vocabulary” as an intermediate representation and used this representation to recognize semantic concepts such as objects. The issue with an image is that there is no obvious way to split it into a set of words. So the researchers proposed the following. Images would first be split into small image patches. These patches would then be grouped, using learning algorithms, into visually consistent groups where each group can be understood as a visual word. Each piece of an image can be mapped to one of these visual words. This visual vocabulary is very simple to learn and yet offers a higher degree of abstraction with respect to the images themselves. Xerox researchers then “showed” computers bags-of-visual-word representations that correspond to different objects to train them to recognize them.

The paper that describes this idea (“Visual categorization with bags of keypoints”) had a tremendous impact and created a paradigm shift in the computer vision research community. Researchers moved away from small scale problems studied in laboratory settings, and started to address much larger scale realistic problems. The “bag-of-visual-words” model is now a de facto standard in the research community. Almost 10 years after publication, it remains one of the most cited articles in computer vision. The vast majority of algorithms proposed since then build on the same seminal “visual vocabulary” idea.

This technology has been applied to many problems
of high practical value. In Xerox, it has been used in applicative scenarios as varied as document routing in scanning workflows, vehicle recognition in surveillance videos, product recognition in retail businesses or image aesthetic analysis in communication and marketing.

To get a better feeling of what Xerox visual search engines can do today, see the categorization demo on Open Xerox.

**About the Authors:**

Diane Larlus is a research scientist in the Computer Vision group at Xerox Research Centre Europe. Her main interests are object recognition and localization, and more generally machine learning applied to computer vision.

Florent Perronnin is principal scientist and manager of the Computer Vision team at Xerox Research Centre Europe. His main interests lie in the application of machine learning to computer vision tasks such as image classification, retrieval or segmentation.
A revolution is about to disrupt how news is created and the very nature of journalism. These changes are attributed to automation, analytics and crowdsourcing all of which touch a broad sector of business.

Throughout history, journalism has constantly changed and adapted. It has integrated technological progress in printing presses and copying, and faced the rise of radio and television. On the societal front, it has adapted to reader demands for accurate information and for gossip. It has delivered information from the battlefront, provided political partisanship articles and conducted investigations.

Today it faces a whole new level of upheaval with the advances in computer science: the Internet, natural language generation, machine learning, and interconnected personal devices and social platforms that permit the rapid dissemination of news.

These advances have begun to disrupt how news is created and the nature of journalism. The changes are based on four main trends:

1. Print to online
2. Machine-generated news
3. Data-driven journalism
4. Citizen journalism

Print to Online

This is the oldest of the recent revolutions, and some may even consider it as passé. Yet aside from some notable exceptions, media outlets are still struggling to adapt their revenue strategy to the fact that most readers consume their news online instead of buying a print newspaper. Moreover, they expect free online access. Solutions for providers have emerged in the shape of subscriptions or payment for access to full articles, online advertising and other strategies.

Machine-Generated News

Recent advances in artificial intelligence and the diffusion of tools that leverage them are accelerating the automation of certain white-collar jobs that were traditionally considered exclusive to humans. Journalism is highlighted as one of the first categories of jobs that could be lost, spurred by advances in natural language generation and fast data processing of structured data.

Some startups have already started to create news summaries from feeds of structured data. The application of these systems is for now restricted to the production of summaries with very limited variability in restricted topics such as sport or finance. When Kristian Hammond, a co-founder of one of these start-ups, claims that in 15 years, 90 per cent of news articles will be written purely by software, he is saying more about the growth of this activity than about the reduction of articles written by professional journalists. It will certainly
automate some low-investigative journalist tasks, but this will create a new market not disrupt an existing one.

Data Journalism

The term «data journalism» may sound like a pleonasm, because journalism is arguably all about transforming data into a story. The «Data Journalism Handbook» defines it as the combination of traditional journalism (notably the nose for news and ability to tell a story) with «the sheer scale and range of digital information now available». It is the growing availability of digital (and structured) data, together with the scalability of data-crunching algorithms and advancements in data visualization that defines this new trend.

A reason for the current popularity of the term may be the reader’s higher expectations of news outlets. They have become used to Internet companies and portals providing them with easy to navigate interfaces and visualization tools with which they can dive into the data. An increasing number of governments now even mandate that certain public information and data (belonging to or held by the government) be made publicly accessible. Alternative associations like WikiLeaks also provide a steady flow of raw data. Combining access to such a vast realm of data with reader expectations has resulted in several news providers applying techniques to extract information, summarize and visualize the data. The Guardian, The New York Times and Der Spiegel are some of the established outlets spearheading this trend, and on whose websites the reader is witness to real cases of data-driven journalism.

The algorithms which mine news and raw databases and convert them into mathematical objects are being combined with other algorithms that read these objects to output news stories and summaries. From this combination is arising a new business niche, the one of personalization. Its goal is to give real-time summaries of very specific information as soon as it appears. One day these may be used by all news consumers but for now most of the target users are specific decision-makers for example those in financial markets.

Citizen Journalism

This last trend may well be the most revolutionary. Still in its early stages, its real impact is more difficult to predict. The idea behind citizen journalism is that anyone can comment on, report news and even be the first to break it. A journalist does not necessarily have to be at the right place at the right time. It may be enough that he or she is somehow getting the flow of reports of many people and aggregating, filtering and adding background and analysis to a story. This «somehow» is made possible with the ubiquitous presence of personal devices that have access to the Internet and a camera. A presence which will increase in magnitude as wearable devices – like digital glasses or smart watches - become mainstream.

The second necessary ingredient for citizen journalism to work is a platform to share citizen reports. Micro-blogging platforms (most notably Twitter) are fulfilling this role and much was claimed about the importance of Twitter in the so-called Arab Spring of 2012. The truth is that, in unexpected events like the Woolwich shooting in the UK or the Boston marathon bombing, where very few facts were known, journalists responded to their readers' thirst for information by swimming through the sea of tweets and reporting their findings.

The real-time nature and public demand of this approach exacerbates the historic tension of broadcasting breaking news before verification. This lack of trust is reflected in a recent survey, which showed that, irrespective of the source of a story, 60 per cent of people turned to an established outlet for confirmation of what was being reported.

The need to discover, filter and deliver information is a wonderful playing field for algorithms that organically incorporate incertitude through machine learning. The UK-based start-up storyful.com adds to this the importance of having trustworthy people who can
confirm or deny a given piece of information. Their slogan «news from noise» encapsulates the problem and challenge of this new kind of journalism.

Reaching Beyond Journalism

Machine-generated content, data-guided decision making and harnessing the wisdom of the crowd is radically changing how news is created and published. But these trends are not confined to news. Their impact can be witnessed across a broad range of fields under the general terms of automation, analytics and crowdsourcing. Customer care is but one example of an industry where automation is used but, at the moment, is mainly restricted to IVR’s (Interactive Voice Responses) and instant messaging. Analytics are becoming widely adopted in this field to detect trending issues by mining customer comments. Much less mature are emerging applications of crowdsourcing to harness and share the knowledge published in specialized Internet forums.

As the face of journalism changes and embraces new technological opportunities, many other document-intensive business processes have a lot to learn from this transformation.

About the Author:

Matthias Gallé is a research scientist in machine learning techniques applied to document access tasks at the Xerox Research Centre Europe. His main research interests include machine learning, grammatical inference and text algorithms.
Can “Made Up” Languages Help Computers Translate Real Ones?

In the late 1880s a Russian-born doctor Leyzer Leyvi Zamengov created an easy-to-learn, politically neutral language aimed at transcending nationality and fostering peace between people with different languages.

Dubbed “Esperanto”, the artificial language still has a significant presence in over 100 countries today with estimates of fluent Esperanto speakers ranging from 10,000 to 2 million worldwide.1

The idea of communication via artificial language by Zamengov and others, today is known as constructed languages or “conlangs”. Hollywood and the film industry in general is particularly fond of conlangs, which offer a quick way of instilling a certain kind of exoticism in even the most run of the mill scenarios. What do the box office hits Star Trek, Game of thrones and Avatar have in common? They all have their own conlang: the Klingon, Dothraki or Na’vi languages, which were specifically designed to enrich their different universes with a touch of the strangely unknown. The sounds, vowels and consonants of these languages were chosen exactly as you would choose costumes and makeup to convey the brutality or the softness of these imaginary worlds. These constructed languages are in fact languages in the proper sense of the term, in which dialogues are written, to which linguistics theories apply, albeit without the complexity of our natural languages, since these languages can be described in extenso. This inherent factor means that conlangs are also an efficient way to help computers translate text into multiple languages – on the fly.

The Challenges of Machine Learning

Despite 60 years of research in computational linguistics, computers still fail to grasp the meaning of the most simple of sentences and ambiguity still plagues the most refined linguistic theories. We human beings solve ambiguities with our understanding of the world, with our capacity to put utterances back into their context. In contrast, the most advanced techniques such as machine learning, can only count words or phrases, relying on co-occurrences and word distances to make sense of texts and documents.

For instance, a very simple sentence such as “The dogs are loud” will be correctly translated into French, by a well known translation site, as “Les chiens sont bruyants”. However, if the sentence is modified into “The dogs are way too loud”, then the same site yields “Les chiens sont beaucoup trop fort”, which means “The dogs are way too strong”: a quite surprising semantic drift from the

1 http://en.wikipedia.org/wiki/Esperanto
original English sentence. Furthermore, the agreement between “fort” and “chiens” is lost in translation. Thus, even the most advanced translation systems can be easily disrupted with a few modifications. This is because texts are ambiguous, terribly ambiguous. Natural languages have evolved in a rather organic way, without any actual plan, hence this inextricable fabric, with which our computer programs have so many difficulties. At each step, syntactic analyzers or parsers are confronted with ambiguous words and ambiguous constructions. The combinatorial aspect of natural language is such that a computer program might end up with thousands of potential analyses for a regular sentence of twenty words. Machine learning techniques have tried to reduce this complexity, weighting words and constructions to feed complex classifiers, but ambiguity cannot always be reduced to correlation. Parsers are dumb, they analyze sentences one after the other without keeping track of past analyses. What we need is a non-ambiguous representation which could be used not only to store previous analyses, but also data from the real world, an intermediary structure which would be close to a human language but would have the properties of a computer language: a constructed natural language that could be compiled as a program, in other words a conlang. John McCarthy, the man who coined the term “Artificial Intelligence”, had this idea back in 1976 when he proposed to solve Natural Language Processing issues with what he called “Artificial Natural Languages”, another name for “conlang”.

Lingvata is one of these languages, designed to be free of all forms of ambiguity. An artificial language that can be used as an intermediary step for computers to translate, for instance, from one language into any other. Lingvata uses suffixes, which encompass one single part of speech, to avoid category ambiguities as in “drink” which can be either a noun or a verb. Lingvata provides a unique ending for nouns, pronouns, adjectives, verbs, prepositions, determiners and adverbs, eliminating the risk of ambiguous interpretations and errors. Words are simply created by combining a semantic root with one of these suffixes.

For instance, the root “parole” is related to speech.
- paroleta - with the noun suffix “ta” means “speech”
- paroleiag - with the verb suffix “iag” means “to speak”.

Thanks to this simple mechanism, Lingvata can be enriched with as many words as necessary, without introducing any homonyms or too many synonyms.

Lingvata also provides a mechanism to avoid syntactic ambiguity, based on Latin as a model. In Latin, the role of the different words in a sentence is governed by their suffixes or case markers. For instance, the sentence: “domina rosam amat”, means “the lady loves the rose.” The termination “am” indicates which element in the sentence is the direct object. If you shuffle the suffixes, you also change the sense as in “dominam rosa amat”, which now means: “the rose loves the lady.” This is a very efficient way to encode syntax, as each case marker conveys only one possible syntactic interpretation. Thus, Lingvata has case markers to indicate not only a direct object or an indirect object as in Latin, but also specific combinations to encode verb and noun complements. These follow a strict word order to make the whole syntactic process totally deterministic, for instance, the verb is always at the end of the sentence. Lingvata offers four different terminations for case marking (vs. six in Latin), which are shared by all categories:
- Nominative: subject, no specific ending. Example: “paroleta”
- Accusative: direct object. Ending is always ‘n’. Example: “paroleta n”
- Genitive: Noun complement. Ending is always ‘s’. Example: “paroleta s”

The Esperanto language was the source of many of the semantic roots that we use in Lingvata. Since most of these roots have already been translated in many natural languages, it proved the most efficient way to bootstrap our own implementation of the Lingvata language.

Word order, case markers and terminations in Lingvata are of course arbitrary. One could design a completely different grammar that would still retain the same properties. However, if designing a conlang is actually fun, it requires quite a lot of work and experiments to achieve the right balance between simplicity, conciseness and expressiveness.
• Dative: Prepositional phrase. Ending is always ‘d’
  Example: “parolletad”

Our previous Latin example “the lady loves the rose“ would then translate as: “Dameta rosetan ameiag “.

The grammar also provides mechanisms to handle clauses and conjunctions, but most of the sentences rely on the few above rules. As an experiment, we wrote a little text in Lingvata and checked if we could automatically translate it into French and English.

Thanks to the simplicity of the grammar, the analysis of a Lingvata sentence is straightforward. It requires less than 50 rules to cover all aspects of the language, which enables us to translate each sentence into French or into English in less than a few milliseconds on a basic computer. In comparison, the English grammar comprises more than 3000 rules. We have also developed a tool that takes as input a sentence in French and translates it into Lingvata. We can then fix the errors in the Lingvata output, since we know that the French analysis is not always reliable, and store the results in a file, which can then be used to translate into any language for which we have a generator. This could be used for example for website content in multiple languages where the original text would be in Lingvata. The system would then translate the content on the fly into the user’s language, removing the necessity to maintain as many versions of the text as there are languages.

In a certain way, the most compact way to store the semantic representation of a text is...the text itself. For a long time, linguists have tried to formalize languages into strict mathematical frameworks, but languages have proved to be so elusive that most theories “leak” – that may be why we call them natural. Machine learning techniques, despite their careful injection of hard science into the problem, did bring some improvement, but the best systems still fail to provide a precise and reliable analysis for too many cases. Today, with the advent of the internet, textual information is everywhere. Yet, the ambiguity and complexity of natural languages makes it quite difficult to draw on these resources in an efficient manner. On the contrary, an Artificial Natural Language representation keeps the whole spectrum of linguistic data intact with very little or no loss of information.

A paragraph or a sentence written in a conlang is a description as precise as any piece of text and at the same time the semantic encoding of that text: a symbolic representation which sits half-way between man and machine.

About the Author:

Claude Roux received his Ph.D. in syntactic parsing algorithms from the Université de Montréal (Canada) in 1996. This work was the basis of the Xerox Incremental Parser which can be accessed from our virtual lab Open Xerox. His main interest lies in syntactic parsing and formal language theories. He is the creator of the Lingvata conlang.
Do you Act Like your Computer (Designer) Thinks you Ought to?

Since everyone from office worker to police officer uses computers nowadays, their design has a major impact on our life. To make systems design successful requires understanding exactly how work is carried out, including the often unnoticed human expertise used every day in what we call unskilled jobs.

Ethnography uses systematic data capture and rigorous analysis to uncover what people actually do instead of just what they say they do. Xerox, a pioneer of applying this research to technology innovation, uses ethnography to identify unmet needs, recommend practices that work and avoid design constraints that impede how work is done.

When carried out prior to design, ethnographic study reveals the real contingencies of the work and helps us design systems which can handle that complexity; post-design ethnography can make the often implicit and taken-for-granted models underlying systems design visible.

In her ground-breaking ethnographic study, “Plans and Situated Action”, Suchman1, demonstrated what happens when design bears little resemblance to how people act. In the study even engineers had major problems using early photocopiers because system design embodied cognitive principles about how people plan and act, which bore little relation to how users actually interacted with these machines.

In another study, Whalen et al2 provide an eloquent dismissal of the founding principles underlying the implementation of computerized “expert systems” in a call centre. These systems were intended to allow unskilled call agents to answer customer problems in the place of trained experts. However, the so-called expert systems were not able to respond to the contingencies of the interaction in the way human experts could, leaving the untrained agents with little option but to escalate their calls. Most of us have had at least one customer service interaction which ended in frustration because the representative seemed unable or unwilling to deal with our request. Typically this is caused by poorly designed workflow systems rather than truly unhelpful people.

Human Skill Versus Computing Power

Whilst automation and the pursuit of cost reduction are not inherently bad, the danger comes from an oversimplified model of low-skilled work: just because workers do not need advanced education or long apprenticeships, does not mean a computer can do that work better. There is a fundamental difference between the human skill and computing power: we’re good at interpretation, improvisation and interaction, computers

---

are good at large scale number crunching.

If ethnographers ruled the world of work it would be a very different place. Automation of manual tasks that computers can do just as well or better than humans makes sense. It’s important to recognize however, that the main use for computers should be to enable people to capitalise on their expertise - whether it be highly refined professional knowledge or mundane, everyday sense-making. We can do this through designing systems which 1) enable people to use their skills and reasoning to deal with the problems which routinely crop up, rather than restricting them to rigid workflows or 2) consist of unique combinations of human and digital input to create output that neither can complete on their own, e.g., crowdsourcing image categorisation, where people identify the semantic parts of an object (e.g. wing, beak, etc. of a bird) and the machine applies the rules to identify exactly which object this is (e.g. this bird is a speckled hummingbird). Call centres provide a good illustration of what can go wrong and right with computer system design.

Agency and Expertise

Time and again, ethnographic studies have shown the importance of agency (the capacity of a human to act). These studies have highlighted the myriad human skills which go into even mundane work. For example, we compared two parallel government processes, each consisting of a call centre and a processing centre. Where the teams were collocated, the workers could circumnavigate the rigid workflow with face-to-face communication to ensure vulnerable citizens got the benefits they were due without delay. In another setting the teams were distributed, had no agency to step outside of the workflow and could only communicate using official channels. In this situation, the call centre could not answer many queries - resulting in caller frustration - and it was necessary to employ someone full-time to follow up on citizens’ complaints, most of which could have been avoided.

In another example, we worked with a customer service company that wondered why customers chose to call the call centre when there were online resources to help solve customer problems. By putting the same resources used by the call centre agents online, the company was hoping customers would choose to solve problems themselves, making it possible to remove the middleman.

Studying the call centre agents as well as customers using the online resources we identified all the extra work that the agents did to solve the customers problem – from persuading unwilling customers to troubleshoot in the first place, to translating the customer’s language into that of the online system.

When customers used the online system alone we found that even when a search for answers returned the right results, customers often did not realize it was the right result. Call centre agents, on the other hand, were skilled in making the same search result relevant to the customers’ problem. They used their semantic skills to guide the customer through the problem-solving process. The company realized that simply removing the agents was not an option.

These examples point to an important two-fold lesson: even low skilled workers are often engaging in semantic work which is not apparent at first glance but which is almost impossible for systems to replicate. Secondly, tight control through rigid workflows is rarely optimum, as it tends to stop workers from using their skills effectively. Design should support and enhance human expertise, rather than attempting to automate and control it.

About the Author:

Jacki O’Neill is a principal scientist at Xerox Research Centre Europe. Her main area of interest is in the design of useful, usable and innovative computer systems, through both the detailed understanding of work practices and a consideration of the interaction of the social and the technical in prototyping and development work.
Will Translation Ever be Fully Automated?

Translation requires a deep understanding of language yet this ability lies far away on the horizon of machine translation. Today’s statistical methods are however moving in the right direction, providing imperfect yet useful translations. Short-term progress is likely to come from stronger statistical learning combined with linguistic knowledge and better modelling of the target language.

Since its beginnings in the 1940s, machine translation has been a subject of fascination for computer scientists. It has a unique position among artificial intelligence problems because it is both a challenging intellectual task, like text or scene understanding, but also - in contrast to those tasks - one that produces overt traces, the translated texts, which are relatively transparent to us human judges.

In the early 1990s, the field experienced a revolution. To the astonishment of the machine translation (MT) establishment, a team of researchers at IBM demonstrated that an MT system could be automatically learnt from a large set of text using statistical techniques, instead of requiring the delicate hand-crafting of linguistic rules common at the time.

Since this initial experiment, Statistical Machine Translation (SMT) has contributed to a significant resurgence of interest in machine translation, and led to popular generic SMT services such as Google Translate, which about 200 million people use 1 billion times each day.

Despite the huge progress experienced over the last 20 years, machine translation still falls far short of producing results of comparable quality to human translation. So, will translation ever be fully automated?

Being old enough in the field to have participated in the development of both rule-based and statistical translation systems let me share some brief thoughts about this slightly chimerical question.

Machine Translation and AI-Completeness

In the field of artificial intelligence, ‘AI-complete’ problems refer to challenges that can only be solved by programs that perform at the level of human intelligence.

It is not difficult to illustrate the AI-completeness of translation. Take for example the French text: “Le policier avait poussé la femme dans le canal. Sa condamnation fut rapide” which could translate into English as “The policeman had pushed the woman into the canal. His/Her conviction was quick.” Because the agreement of possessive adjectives in French is with the following noun, but in English with the owner, the only way that a translator can determine whether “sa” should be

translated by “his” or by “her” is through a complex process of reconstructing implicit facts not directly mentioned in the text: the woman drowned and died, the policeman’s act is a crime, crimes are judged in courts, etc.

The opposition implicit/explicit represents a major challenge for machine translation. This is because languages are very different in what they mark explicitly and what they leave in the background.

The consequences of this for MT may be illustrated by the following optical analogy:

In the picture, French and English correspond to two different “light sources” that project “shadows” of the same mental representation onto a textual plane. These shadows have many things in common, but also significant differences. These differences correspond to the different aspects of the representation that the two languages make explicit. To translate from English to French, one needs to produce the French shadow from the English one.

The natural translation process requires first reconstructing the common situation using knowledge of the English light source along with commonsense understanding. It then projects this onto the plane using the French light source.

It is however often possible to produce good approximations of this complex process by focussing on regularities that can be observed at the level of the textual plane. This has been the route followed by most rule-based and statistical machine translation systems.

Here, SMT has some decisive advantages over traditional rule-based systems:

It is unrealistic to hand write rules covering the myriad idiosyncratic constructions that need to be translated case-by-case between different languages. In contrast, SMT systems are able to automatically learn many such constructions from bilingual corpora; thus Google Translate has no difficulty translating Tom “is hungry” into Tom “a faim” [Tom has hunger] despite the distinct constructions, and this observation holds for many similar expressions.

• An equally important advantage is that SMT systems embed a powerful mechanism which is a probabilistic language model of the target language. Such a model provides an estimate of how likely a certain sentence in the target language is, based on generalisations from observations of a large number of existing target language texts. Such estimates can often provide a “proxy” to simple commonsense inferences, helping to disambiguate between different choices; thus the Google service translates “le temps est souvent pluvieux” into “the weather is often rainy”, but also “le temps est souvent limité” into “time is often limited”, correctly selecting the right translation for the ambiguous French word “temps” in both examples.

Current SMT systems are however still limited in their capability to generalize. For example, slightly enlarging these examples tends to produce wrong translations: Tom is extremely hungry ► Tom “est extrêmement faim”, le temps “est souvent désagréablement pluvieux” ► time is “often unpleasantly wet”. Here the target language model is “misled” by the increased distance between semantically clashing words (“est” with “faim”, “time” with “wet”), and is not able to detect the inconsistencies.

2 The World Atlas of Language Structures (WALS) http://wals.info/ provides an extended compendium of the many such typological differences; some languages do not mark Tense or Gender, some others obligatorily indicate things like “Evidentiality” (how the speaker came to know what he says), etc.

3 Figure inspired from http://www.researchgate.net/publication/245023525.
Solutions on the Horizon

Returning to our original question, if we wished to focus on the AI-completeness argument, our answer would have to be something like: “Fully automated translation will probably happen in the future, but around the same time at which our streets will be populated with the kinds of androids depicted in the recent Swedish sci-fi series Real Humans”. However, many other tasks that are already highly automated are, or can be construed to be AI-complete, e.g., speech recognition or even optical character recognition. While MT is more AI intensive than these tasks, much progress can be made by the MT community without having to wait for a full solution to the general AI problem. This can be done by focussing on solving many specific difficulties of translation which currently are poorly handled but have potential solutions on the horizon.

Let me outline one possible direction for such progress. Professional human translators typically translate from a foreign language into their native language, where they are much more comfortable with syntax and linguistic nuances. In fact, even novice learners of a foreign language can often do a good job of translation with the help of a bilingual dictionary as long as their linguistic competence in their native target language is high. To some extent, the understanding of the source text may itself often be assessed by checking that the translation makes sense. We could say that these people are exploiting their ‘Natural Language Generation’ capabilities in their native language.

These observations of human translators are interesting to relate to the current status of SMT. While it is true that target language models play an important role, most effort is still devoted to acquiring bilingual corpora. This can be difficult, especially for domains which are not well represented by previous translations. Monolingual data in the target language, for most domains of interest, are much easier to find. It might therefore make sense to focus more attention on better exploiting the monolingual data to produce deeper generation models for the target language. Learning richer language models from these data could prevent frequent errors of the kind illustrated above (Tom “est extrêmement faim”, “time is often unpleasantly wet”). Other common generation errors may be more difficult to handle based on data only and without injecting prior knowledge of the grammatical rules of the language, including some intricate prescriptive rules that we learnt at school.

If translation is ever fully automated it is likely that rich statistical models emphasizing the target language combined with classical linguistic knowledge will be major ingredients, and the good news is that these are dimensions on which we can already make progress today.

About the Author:

Marc Dymetman is a principal scientist at Xerox Research Centre Europe. His main research interests are statistical modelling of natural language, machine translation and machine learning.
How Economics and Machine Learning Can Tackle Transportation Congestion

In an ideal world, when you want to travel from point A to point B, it’s just you, your horse and the empty prairie. When you need to park, you just tie your horse to the nearest saloon and you’re done. However, as we all know, we do not live in an ideal world. And when we travel from A to B it’s not that simple. It’s you and your car… and a whole lot of other cars.

The description just given is a common case of demand exceeding supply. It results in traffic jams and drivers circling blocks hunting for a vacant parking space. One way to solve this problem is by extending the supply. Governments deliberate and decide whether tax payers’ demand for new roads justifies the spending of tax money to build them. Ever since the introduction of cars, extending supply has been the predominant approach to solving this problem.

Yet demand is ever increasing and constructing new infrastructures is costly in terms of the direct costs of building garages and highways, and the indirect costs of reducing the amount of land available for housing and offices. There are also costs associated with environmental impact. So when infrastructure becomes congested it is worthwhile looking at alternative solutions such as a more efficient use of existing infrastructure commonly referred to as “demand management”.

Xerox researchers are developing algorithms that lie at the heart of demand management solutions. This article provides an introduction to the goals of demand management and offers a glimpse under the hood of a project to reduce parking congestion in the city of Los Angeles.

Demand Management and the Economics of Transportation

The basic argument for demand management solutions in transportation dates back to at least the middle of the 20th century. Yet, because of technological and political difficulties, its use on the scale of a city has appeared only recently.

William Vickrey (Nobel laureate, 1996) provided a famous hour glass metaphor for traffic jams. In a simplified scenario, commuters start at one end of an hour glass on a road that takes them to the other end. At its narrowest part, this road has a certain maximum capacity which determines the rate of outflow. So when a large number of commuters join the traffic at 8:30 a.m. aiming to arrive at work at 9 a.m., not all of them will succeed. The bottleneck will cause a jam and some cars will have to wait. The outflow will be dictated by the bottleneck.

Let us examine the cars’ arrival times. Some will arrive at 8:45 a.m., some at 8:46 a.m., some at 9 a.m. and others at 9:10 a.m. The majority of them will have spent
a significant time in the traffic jam. Now consider a different world: one where the start time of each car is controlled. Suppose the journey without the jam takes 15 minutes. We would ask the car that arrives at 8:45 a.m. to leave at 8:30 a.m., the car that arrives at 8:46 a.m. to leave at 8:31 a.m. and so on. The result is that all cars arrive at exactly the same time as they did before, but now all of them only spend 15 minutes in free-flow traffic instead of the average of say 30 minutes because of the traffic jam. If you work with a value-of-time saved of $10 an hour, every person in the jam gains $2.50 per day.

The reason that it doesn’t already happen like this is that a majority of commuters would prefer to be in the car that leaves at 8:45 a.m. and arrives at 9 a.m. There also is no formal mechanism for drivers to coordinate their departure times. Vickrey argued that the current dynamics of traffic jams like these are an equilibrium state where people balance the cost of leaving too early and the risk of arriving too late.

The idea of demand management is to help drivers coordinate using time-differentiated tolling e.g., the closer to 8:45 a.m., the higher the toll. This would reduce the jam in the bottleneck in a way that ensures that the drivers who are least inconvenienced by leaving slightly earlier or later (or taking public transport) would be the ones to do so. This is the motivation for congestion charges such as those used in Stockholm and Singapore and the smart toll lanes around Los Angeles and Miami.

To conclude this stylized example demand management aims at a more efficient use, instead of an expansion of infrastructure. Drivers are encouraged to avoid peak hours so that more cars can pass in free-flow conditions from A to B.

The reasoning is in principle not restricted to highway access but is valid for all publicly-owned utilities. The principles would apply equally well to parking and public transport. The fact that the infrastructure is paid for by public money and in a sense is owned by everybody doesn’t mean that it is put to best use if it is accessible for free.

### Analytics and Communication

Even though the principles are elegant and could provide a very effective solution to important problems in most people’s daily lives, it is also easy to see why they weren’t implemented directly in the 1950s. First, it is difficult to explain to drivers that they need to start paying directly for a service that was previously paid for by general tax money. The increase in efficiency can be hard to explain and in the original subsidized situation, some users who use the service significantly more than the average tax payer would pay more in a demand based system. These users will be very opposed to any proposed change.

Secondly, it is far from straightforward to predict how many cars will access a particular bottleneck, what the right rates need to be to incentivize drivers to avoid a potential traffic jam, and to notify drivers in a convenient and timely way. This was particularly true given the state of technology in the 1950s. However with today’s real-time sensing capabilities plus smartphone and satellite navigation communication possibilities, the prerequisites for effective solutions are available. Several large-scale projects already are in use. For example, Stockholm has time differentiated tolls to access the city centre; Singapore uses gantries to have time- and location-differentiated tolls, and San Francisco and Los Angeles use on-street parking sensors to let parking rates be guided by demand.

### Xerox and Demand Management for Parking

The LA parking solution, LA ExpressPark, is implemented by Xerox for the Los Angeles Department of Transportation (LA DOT). The algorithms that harbour the analytics and economic trade-offs are designed by researchers at the Xerox Research Centre Europe in close collaboration with LA officials.

Since May 2013 all on-street parking spaces in downtown LA have been equipped with sensors. This gives real
time information if a car is parked in a space or not. The data can be used to guide people to empty spaces using smartphone applications. In addition, machine learning methods can be used to understand and predict demand. In turn, pricing algorithms can use these models to adjust rates to demand.

During the project, algorithms are used to revise rates at roughly three month intervals. The Holy Grail is to iteratively find rates that make sure demand is so well spread out that every side of a block in LA’s grid is nearly always close to being full but has one or two spaces free. That way the spaces see good use and drivers can leave home, make a decision where to park, and are almost guaranteed not to have to circle round to find a free spot.

The project started in June 2012 and, with more than a year’s worth of experience, the detailed impact analysis will commence in 2013. From preliminary results we already know that the rates went down in more places than where they went up and that the average parking rate went down overall. Nevertheless, since the rates increased in very congested areas and decreased in the underutilized areas there is an overall slight increase in revenue. Most importantly, the desired state where blocks were neither underutilized nor congested, increased by 9 per cent.

**Outlook**

Xerox has a long history of innovations that have taken some of the dread or drudgery out of our daily lives. Consider the need for human copy typists before the invention of the copier or the difficulty of correcting an error in a piece of text written with a typewriter instead of a personal computer.

Advances of analytics in transportation have only just begun, so it remains to be seen if these innovations can have similar effects. But be assured we’re working hard to try to take some of the pain out of the daily commute and regular hunt for a parking space.

**About the Authors:**

Eduardo Cardenas is functional consultant analyst at Xerox Services Austin. His main interests include application of machine learning, data mining and data analysis to parking and transportation problems.

Christopher R. Dance is a Xerox research fellow at Xerox Research Centre Europe where he currently manages the Machine Learning for Services research area. His research interests include computer vision, machine learning and mechanism design.

Stéphane Clinchant is a research scientist at Xerox Research Centre Europe. His main interests are natural language processing, information retrieval and machine learning applied to transportation problems.

Onno Zoeter is a senior scientist at the Xerox Research Centre Europe. His research interests are in the overlap of machine learning and economics. He currently leads a project on demand management solutions. The algorithms developed in this project allow parking rates to be better adjusted to demand.
Do Computers Need Grammar Books too?

The world’s first known grammar “book” was conceived by Pānini in fourth century B.C. India, when only a tiny minority of people could read or write, and when their greatest challenge was bare survival. His grammar contained 3,959 rules of the Sanskrit language, which describe how words are composed, how they combine into sentences, and what they mean.

This grammar probably didn’t influence the everyday life of the population as its primary goal was to teach the proper use of Sanskrit for the writers of sacred books. Who could have foreseen at the time that this knowledge of the composition of words, their combination into sentences and the description of their meaning would one day become so critical in today’s world of information overload?

Today, one of the greatest challenges of the “information society” and “knowledge economy” is making sense of and getting the most benefit out of “big data”, 80 per cent of which is textual. Company analysts trying to follow which company bought which company, for example, might be faced with reviewing hundreds of thousands of documents. Doctors require lists of patients eligible for a clinical trial based on hundreds of thousands of patient reports. Companies need to learn about problems with their products which might be reported in hundreds of thousands of forums.

Understanding all the subtleties and complexities of human languages is, and probably will always be, the privilege of the human mind, but given the enormous quantities of texts, man must rely on the help of automated processing. The better automated processes manage to decode the meaning of the texts, the more they can help by extracting useful information from them, and thus mining knowledge.

Integrating grammar in the design of automated language processing tools can be of real help since texts are composed of words in different forms and roles, and those words make up sentences and the sentences convey complex meanings – which are described by grammar rules.
How Does Grammar Help Access Information and Knowledge?

Today, the most widely used tools that help us access information and knowledge in texts are search engines based on keyword search. When you write a word in the keyword box of a search engine you will get a list of documents that contain that word. But could a business analyst ask a standard keyword-based search engine to provide a ‘list of the company transactions’? Could a doctor ask for the ‘list of patients who are eligible for a clinical trial’? Could a company manager obtain a ‘list of complaints’? The answer is clearly “no”.

To illustrate the limitations of keyword-based search for complex queries, let’s take the example of a business analyst who would like to submit the query ‘buyers and companies bought’. Why, after submitting such a query would the title of this news article ‘Microsoft Acquires Sun Microsystems’ not be returned?

In order to be able to return this answer, the search-engine would have to be aware of how words are composed, how they combine into sentences and what they mean i.e., it would have to “know” some grammar. It would need to master at least the following concepts:

- The concept of a company – to be able to match “Microsoft” and “Sun Microsystems” as company names
- The concept of a transaction – to match “acquires” with a transaction
- The concepts of “buyer” and “thing bought” and their expression in sentences – so that it can match ‘Microsoft’ as a buyer and ‘Sun Microsystems’ as a company bought.

But keyword-based search engines are not aware of these concepts. More sophisticated search engines are required.

Beyond Keyword Search

Since the foundation of Xerox Research Centre Europe in the early nineties, one of our main research topics has been natural language processing. Over the past 10 years, the focus has been on information extraction so that automated tools can return answers to more in-depth queries. We have developed what we call ‘FactSpotter’, a sophisticated information extraction tool that takes into account the complexity of language structure, and can navigate the three concepts described in the example above.

Based on linguistic rules, FactSpotter can detect the names of people, companies or locations, dates and various other so-called “named entities ” in texts (and it can do this is several different languages). In the example above, FactSpotter would identify ‘Microsoft’ and ‘Sun Microsystems’ as company names. It would analyse word forms, and provide a formalism, which allows users to constitute lists of words and expressions that convey the same concept. This ability makes it possible, for example, to associate the word ‘acquires’ with the concept of transaction.

FactSpotter can also conduct syntactic and semantic analysis. Syntactic analysis identifies ‘Microsoft’ as the subject of ‘acquires’ and ‘Sun Microsystems’ as the direct object of ‘acquires’. The semantic analysis maps these syntactic functions into semantic roles. Thus ‘Microsoft’ can be recognized as a buyer and ‘Sun Microsystems’ as a company bought.

Besides these features FactSpotter also integrates other
aspects of the complexity of human languages. It takes the context into account to differentiate among several meanings of the same word (disambiguation), e.g. it knows that in the sentence “I can see, and I see a can”, the first can is a verb, and the second is a noun. It also can recognize particular linguistic structures that carry the same meaning, e.g., “buy”, “acquire” and “become the new owner of”, and it has the capability of recognizing different expressions that refer to the same entity e.g., “Microsoft Corporation” and “it” in the sentence “Microsoft Corporation announced after the close today that it will buy Sun Microsystems”. As you can see, these skills are necessary to understand linguistic meaning - something keyword-based search engines simply can’t do.

FactSpotter has been used in numerous information extraction tasks in different domains and languages: in clinical decision making, event extraction and the establishment of chronological order in news articles, the detection of political risk, mining clients’ complaints for customer relationship management, the extraction of biological knowledge from research articles, etc. We are currently engaged in new research that will make it very easy for FactSpotter to adapt to new tasks.

Grammatical rules were created some 2500 years ago and have been taught ever since schools exist - to the regret of many a pupil! From guidelines in writing sacred texts, to regulating national languages, to learning and translating foreign languages, the practical uses of grammar rules have increased over the centuries. Today even computers are more effective if they have been through grammar school!

**About the Author:**

Ágnes Sándor is a researcher in natural language processing at the Xerox Research Centre Europe. Her research areas are information extraction from biological and medical documents, news articles and textual enterprise data, rhetorical analysis of argumentative discourse and pragmatic analysis of social media postings.
Monks & Markup

Surprisingly, enriching text with tags dates back to the sixth century.

Markup is about enriching text with tags so computers can automatically process written words. XML and HTML are typically what come to mind when we think of today’s markup languages, but in fact, the first markup appeared hundreds of years ago.

Irish monks in the sixth century A.D. who were unfamiliar with the Latin language from the continent first introduced one of the most important tags still in use today – spaces between words. Standardized punctuation quickly followed. Later, during the Middle Ages, a kind of PML (Paragraph Markup Language) was introduced, similar to today’s SGML but back then only opening tags indicated the beginning of a paragraph. It was not until the 15th century that document layout became what we are familiar with today.

So while we think of markup as modern, it originally was created to help humans read, and has since evolved into an essential tool for computer processing. Where this evolution is headed is important, given the growth of documents and their relationship to automated workflows. While the production of documents is not growing at quite the same exponential pace of numerical data, it is still huge and critical to many businesses. Just as the monks relied on tags to help humans read, the effort to “teach” computers how to structure documents will lead to more efficient workflows.

Dividing Content from Form

Markup has a long tradition in print shops where it is used to format and correct manuscripts. This tradition was passed over into photocomposition before computation with languages such as:

<CC 0,5, 12>Nortext,

:i.Ordered GML (Generalized Markup Language)

:bu .b Troff :

These first computed markup languages were more or less a copy of existing manual practices, and corresponded with textual streams where procedural tags about layout were added to explain how content must be laid out.

At the end of the 1960s, under the impetus of the Graphic Communication Association the GenCode (Generic Coding) Committee was created with the goal of reflecting on the separation of document information content from document format.

1 Pause and Effect: An Introduction to the History of Punctuation in the West, Malcolm Beckwith Parkes, University of California Press, 1993

2 http://www.troff.org/
Their conclusions, embodied in the “GenCode® concept”, were that:

- different generic codes were needed for different kinds of documents. In other words, there is no universal document model. Each document type must be described.
- smaller documents could be incorporated as elements of larger ones. Hence the infamous “tree” structure used for representing documents.

Furthermore, markup languages [should] “greatly facilitate the sharing of data and the integration of diverse types of software, yielding a new era of efficiency and flexibility.”

The first technological achievement that put these principles into practice was GML (Generalized Markup Language) designed by C. F. Goldfarb, E. Mosher and R. Lorie in the 1970s. Then followed its offspring: SGML, HTML and XML. XML, which leveraged the experience of SGML but which was easier to manipulate, emerged as an industry standard and is recognised as “the universal format for structured documents and data on the Web.”

Well specified and based on international standards such as Unicode, a set of tools was quickly developed by the community and made available to users. In line with the first GenCode concept of incorporating smaller documents within bigger ones, specific languages were also designed including SMIL for multimedia data, RDF for resources or MathML for mathematical formulae.

Following the second GenCode concept, a structured document is one that is hierarchically organized by elements explicitly marked up with a tag. This markup, which reflects the semantic structure of the document, does not correspond to layout instructions. The layout instructions are given by a style sheet (XSL for instance), which indicates how a tag (e.g., its content) must be laid out.

Finally, a document type called schema, may be associated with a document, so that it can be validated e.g., to check it complies with a specific model. This warranty prevents inconsistency in further processing. This separation of content from form is a benefit for publishers who now target multiple output channels. It is also essential in automated document processing to:

- be able to reuse content to reduce costs in producing new documents
- repurpose content which targets different audiences and channels
- search based on content and markup
- check for compliance which also enables long-term archiving.

Who Uses Markup Today?

It would be reasonable to assume that anyone who is literate can write a digital document using simple word processing software. But who is able to actually design a markup document? To do so one must select the correct document model, properly markup its pieces, validate compliance against its chosen model and select the formatting style sheet to display it. After having worked 20 years on markup, Brian Reid, father of the Scribe system, presented his reflections on markup technologies in 1998 one of which is still valid to this day: Most people won’t use abstract markup even if you threaten them. Document markup is beyond a layman’s skills and is reserved for professionals. Word processing software provides some assistance with spell checkers and grammar checkers. Yet today’s markup checkers are at the level of the very first spell checkers from the 1960s which could only look up a word to see if it existed in a checklist. Unstructured documents accumulate in our PCs every day and the “new era of efficiency and flexibility” that markup could bring is still a dream.

If most of these documents are unstructured, how can we automate their processing? How can they be easily structured to be fed into automated workflows? If

---

4 http://www.w3.org/MarkUp
6 For Reid, “Markup is a mathematical abstraction in the field of data/information.” ibid. Technical Edition
humans are not ready to accept this burden, computers have to be taught to undertake it. Over the years, a team at Xerox Research Centre Europe has been investigating the challenges related to the automation of document processes, including document understanding, document conversion to XML and schema management. The team addresses research challenges relevant to analyzing and understanding document collections based on their layout and structural organization. These methods can be thought of as going beyond what Optical Character Recognition systems do at the character and word level to reconstruct higher level structures, such as document sections, tables of contents, indexes, etc. to automatically markup documents.

The main difficulty lies in the enormous variety of document content layout. To exhaustively inventory all the possibilities is a never ending and expensive task. One alternative is to inverse this pattern recognition problem by relying on constraints that structure the different elements in a given document (for instance the incremental relation between two page numbers), without having to describe their layout characteristics.\footnote{H. Déjean, J.-L. Meunier, Logical document conversion: combining functional and formal knowledge, Proceedings of the 2007 ACM Symposium on Document Engineering, pp. 135 – 143.} We simply learn for each document which layout has been used, and mark it up accordingly. These methods have been successfully applied to large-scale customer cases where required information has been extracted from millions of pages, automatically feeding customer databases. Some examples of this technology are available as online web services on Open Xerox including the popular PDF to ePub converter. This service will automatically convert for you your PDF file into an ePub file, so that reading will be optimal on tablets and readers.

**About the Author:**

Hervé Déjean is a senior scientist at Xerox Research Centre Europe. His research interests lie in structural pattern recognition and his primary objects of study are texts and documents. He is one of the inventors of the abovementioned PDF to ePub converter.
Xerox Living Lab: The Customer’s Role in Our Innovation Process

The creation of a research centre in Europe in 1993 was warmly welcomed by Xerox sales in Europe, which had a burgeoning interest in new technology and desire to meld the reputation of Xerox as an innovator with customers’ needs.

This invaluable link between research and sales has had a great influence on the centre’s growth and success. Created to drive the corporate transition in becoming a services-led technology business, the centre’s focus has always been on targeting innovative software solutions and services. The customer’s role in this mission is reflected in the centre’s “work practices” research competency which is based on ethnography applied to the work environment, and its Technology Showroom, an innovative effort at the very heart of the centre where researchers demonstrate their findings to Xerox customers and collect feedback.

Research and the Customer: Some Lessons Learned

Lesson 1: Customer show and tell

When XRCE opened, every European country was considered as an independent business unit. This organizational structure enabled ownership and the creation of a network of account managers that would directly link customers with research efforts by coordinating and promoting visits to the centre’s Technology Showroom. Some of the early visitors were surprised by the technologies they saw. Software development seemed very different from research related to copiers or production printers. However customers became excited by the idea of more services provided by Xerox. In a business world dominated at this time by paper-based information, Xerox was using its expertise in documents to boost business productivity. As account managers brought the first stream of visitors from all over Europe to the showroom, researchers gained insight into the topics that would interest customers.

At this point, the Technology Showroom was mainly used as a sales and promotion tool that would contribute to enhancing the customer relationship. As the European research centre better understood the value of directly involving customers in its research activities, it began to devise a recipe for getting the most out of customer meetings. Sales teams helped identify customers who were willing to innovate, had a real problem to solve and would support the strategic value of the centre’s innovation efforts by following the requirements of an experiment and deploying the solution.

Lesson 2: Creating good demonstrations, telling good stories and listening to the customer

One key challenge for researchers is communicating the value of a technology to customers. While not everyone is a Steve Jobs, a good demonstration and polished storytelling skills go a long way in helping create excitement and interest. For example, a researcher
working on classification technologies might take the time to share the difficulty he or she had with managing personal holiday pictures. It’s also important to learn as much as possible about the customer, and adapt the demonstration to a customer’s specific work process. Customizing a demonstration by simply adding the logo of the company you are presenting to is another simple way to capture a customer’s attention. This approach is very connected to the notion of Minimum Viable Product advocated by entrepreneur and author Eric Ries which says that as a researcher learns more from his demonstrations, he incorporates new features, leading to a demonstration that is more convincing to investors.

Lesson 3: Establish a methodology and find a name for it

When I discussed with Paul Millier, professor of Markets and Innovation at the Lyon business school (EM Lyon) our effort to involve customers in our innovation process, he encouraged me to continue this path and immediately adhered to the name of ‘Dreaming Sessions’ that we had chosen for our efforts. This was an expression we had coined with Sophie Vandebroek, Xerox CTO, to describe our customer engagement sessions.

Methodology is equally important as it enables a wider use of practices across an organization. Our process starts with the Customer Led Innovation team receiving requests both from the sales organization and researchers to organize a “dreaming session” with customers. The team is in charge of connecting the dots: selecting the right technologies to discuss with the customer depending on their industry, their role and what the researchers would like to learn.

The second stage is a proposal describing the dreaming session which includes an agenda and description of expectations. The third stage is the preparation of the dreaming session: demonstrations have to be prepared with an increased focus on customization to the customer needs. The Customer Led Innovation Team works with researchers to decide on a list of questions that will be answered by customers.

The third stage is the dreaming session itself. It happens when all stakeholders agree on the content. During a session, the Customer Led Innovation team is in charge of creating rich interactions between customers and researchers. Scenarios are established for using all presented technologies. At the end of the Dreaming Session, customers are asked to rank the relative importance of the different technologies. This information is used in order to prioritize which technologies could be pursued with pilots. As part of this stage, notes are captured and sometimes video can record the interactions.

The fourth stage consists in sharing the knowledge acquired during the Dreaming Session: this information can be critical in order to guide the different applications of a technology.

The fifth stage consists in engaging the customer in a pilot, where the technology would be deployed. The opportunity for pilots is the best way for researchers and Xerox to prove the value of a technology and involve the lines of business that will be delivering the new technology or service on a wider scale.

Lesson 4: it takes teamwork

There are two aspects of Dreaming Sessions that are critical: listening and relationship building. Involving researchers has many advantages. It gives them the opportunity to listen and collect feedback. They are skilled and trained in demonstrating the technologies, and the sessions help illustrate the business value proposition of a technology. However, there needs to be a team in charge of facilitating the relationship with the customer during the demonstration.

Lesson 5: Decide on objectives and metrics to improve the activity

As with any activity at Xerox, Dreaming Sessions have
objectives. When the sessions first began, it was not always easy to find the right balance of objectives. From the innovator’s point of view, what should ideally be measured would be the number of ideas, comments and scenarios received when discussing a technology with customers. From a customer relationship standpoint, it would lean toward the additional revenue generated after having invited customers to a dreaming session. Both areas are very difficult to track, so we decided to keep it fairly simple by setting objectives associated with the number of Dreaming Sessions organized per year, the number of researchers involved in Dreaming Sessions and tracking globally the revenue the sessions were associated with and lastly the number of new technologies on display. Setting up objectives against these goals met the needs of all stakeholders.

Lesson 6: Do not mess around with intellectual property

Over the years, Dreaming Sessions have been a point of debate with our lawyer friends because they can potentially generate new ideas. Patent attorneys become nervous when we start to talk of co-ownership. Some rules that we have tried to follow over the years:

• stipulate that both parties should not share confidential information during a dreaming session
• recognize that intellectual property can be created in a way that it is owned by both parties and both parties have the freedom to use it
• discuss ways to jointly file for Intellectual Property.

It’s important to note that things are slightly different if technical intellectual property has to be disclosed. In these cases we’ve worked with thorough and creative lawyers who helped us set the right context for these sessions.

The Future of Dreaming Sessions

What will be the future of Dreaming Sessions? Things have dramatically changed over the past 20 years. I remember one of my first customer events with a European university. We tried to pack in as many technologies as possible in a day, without understanding the whole context and expectations of the customer. The outcome was not very successful, leading to the customer falling asleep after lunch!

Today we know that a lot can be learned when technologies are put in the hands of a customer at a very early stage. New scenarios of applications can pop up, and preconceived ideas can be put aside without spending too much energy on them. Additionally, new modalities have emerged enabled by the digital age. Xerox Research Centre Europe recently launched “Open Xerox”, a test bed where new technologies capable of working on a Web base can be demonstrated and tested. Many of these can be considered individually as Minimum Viable Products, but what about using FaceBook to discuss a technology? This may be possible in the future, but many questions have to be answered. We would have to look for ways to protect the Xerox brand. Competitors, for example, may access the technology and influence its development. Confidentiality might also be difficult to manage. But what an avenue to reach millions of future users!

The work done in the European research labs has certainly influenced the emergence of Living Labs, an open-innovation ecosystem where researchers, future users of the technology and various stakeholders can experiment with new technologies. The future will certainly be more open, participative and will result in better finding applications faster, simply because it has used the best interaction to do so.

About the Author:

Patrick Mazeau is business development manager at Xerox Research Centre Europe. He manages the Customer Led innovation program across the Xerox Innovation Group and thrives on developing the links between research and business.
How Machines Can Learn What Humans Interpret:
Adapting probabilistic topic models to natural language

During the past decade, we have seen an explosion in the amount of digital information. This has led to information overload, making it difficult for humans to make sense of very large document collections, such as emails, digital libraries, news articles or legal documents.

Probabilistic topic models, in part pioneered by Xerox under the trademark Smarter Document Management Technologies, are now routinely used to analyse and explore large sets of documents. Topic models automatically organize documents into semantic clusters or topics, based on the statistical properties of the text that lies within. Their tremendous success (more than 6,500 citations in Google Scholar at the time of writing) can be attributed to their simplicity and appealing interpretation. Yet while successful, current topic models still do not account for the fundamental properties of natural language, which would lead to more diverse and interpretable topics.

What Are Probabilistic Topic Models?

Topic models extract human intelligible topics from text in an unsupervised way. This means that the clusters of documents are automatically learnt from data without any human intervention. Probabilistic topic models posit a generative process for document collections: they propose a probabilistic model (i.e. a set of interdependent random variables), which describe how documents are generated.

To capture the semantics, the key simplifying assumption made in topic models is that documents can be represented by a mixture of topics, which ignore the word order of the text. While rather basic, this simplifying assumption has proven to be effective in practice when one is only interested in extracting the topics. Providing computers with the capability to recognize the topics of a document enables them to identify documents discussing similar content and then suggest these findings to the human user.

The figure below is an example of four topics and a piece of text, where each word is assigned to one of the four topics (arts, budgets, children, education). Each topic is defined by a list of vocabulary words, each one being assigned a probability.

---

A document is then assumed to be generated from these topics as follows:

1. For each document in the collection:
   b. Decide the number of words in the document;
   b. Choose a weight for each topic. This weight corresponds to the prevalence of each topic in the document.

2. Each word in the document is then generated as follows:
   a. Select a topic. The probability of a topic being selected is proportional to the weight it was assigned in 1.b;
   b. Choose a word in the vocabulary according to the selected topic. The probability of a word being selected is proportional to a topic-specific weight.

This probabilistic model not only proposes an appealing generative model of documents, but it also enjoys a relatively simple inference procedure (a collapsed Gibbs sampler to be precise) based on simple word counts, which is able to handle millions of documents in a couple of minutes. Inference is the process of deciding which topics should be associated with the documents. It is done automatically, based on the data-driven evidence. Knowing the topic association is useful in practice as it enables one, for example, to recover documents that share the same set of topics.

### Weaknesses of Standard Probabilistic Topic Models

A practical issue with topic models is the identification of the most likely number of topics describing the data. This is because the identification is a computationally expensive procedure. When modelling real data, the number of topics is expected to grow logarithmically with the size of the corpus. When the number of documents in the corpus increases, it is reasonable to assume that new topics will appear, but that the increase will not be linear with the number of documents; there will be a saturation effect. The issue can be dealt with in a principled way by considering nonparametric Bayesian extensions, a recent trend in probabilistic machine learning.

A second weakness of topic models is their limited expressiveness. The prevalence of a topic in the corpus is correlated with its prevalence in individual documents. Similarly, the prevalence of a word occurring in the corpus is correlated with its prevalence in the individual topics. These are undesirable properties. For example, a good model should be able to identify that a word characterizes a specific topic irrespective of its frequency in the document collection.

Finally, and perhaps most importantly, the probabilistic model postulated by topic models are inappropriate for

---

**Figure 1** (reproduced from reference 1). Topics are defined by a list of words. Each column corresponds to a topic and the words in the list are ranked according to their relevance. The words in the boxed text are coloured according to the topics shown at the top. Each word is modelled as being drawn independently from one of these topics, neglecting the sequential structure of text.

---


modelling real text. Data sampled from the model are statistically distributed differently than real observations. For example, it is well-known that modern languages exhibit power-law properties (see Figure 2).

![Figure 2](image-url)

Figure 2 shows the ordered word frequencies of four benchmark corpora available from http://archive.ics.uci.edu/ml. Let \( f_w \) be the frequency of word \( w \) in the corpus. It can be observed that the ranked word frequencies follow Zipf’s law, which is an example of a power-law distribution: \( p(f_w) \propto f_w^{-(c)} \), where \( c \) is a positive constant. Like many natural phenomena, human languages including English exhibit this property. Intuitively, this means that human languages have a very heavy tail: few words are extremely frequent, while many words are very infrequent. This means that human languages have a very heavy tail: few words are extremely frequent, while many words are very infrequent. This is not accounted for in classical topic models.


The data we observe in practice, such as text, images or social networks, show significant departures from standard distributions encountered in statistics. When our target application is to automatically organize a large set of documents according to topics, we should use models that are able to learn a potentially infinite number of topics and capable of accounting for the power-law characteristics of natural language. Moreover, we would like to increase the model expressiveness, either by allowing more diverse topic distributions, or by favouring more specialized topics, while preserving a simple and efficient inference procedure. This can be achieved by basing topic models on a stochastic process called the Indian Buffet Process (IBP).^{4,5}

The generative model for a document resulting from the IBP-based topic model is similar to that of the standard topic model, except that a small subset of topics is selected before assigning them a weight. Similarly, each topic is defined by a relatively small subset of the vocabulary words, but which follow a power-law. The IBP operates as a binary mask on the discrete distributions defining topics and their association to documents. Topics extracted from the corpus are more specific, possibly assigning a large weight to infrequent, but informative words and they are more discriminative. We observed experimentally that fewer topics were associated to each document.

We currently are exploring how this new type of topic model can be integrated into an Idea Management System (IMS), which can be viewed as a collaborative brainstorming system. In its most simple form, an IMS is a so-called suggestion box, where customers and/or employees can submit feedback or make suggestions for product/service improvements. Large companies such as IBM, Dell, Microsoft, Whirlpool, UBS or Starbucks have deployed such systems to better support innovation with the aim of capturing the collective wisdom residing in the employee and/or customer base. Xerox is adapting IMS to other domains, such as urban planning and policy design, facilitating the communication between citizens and political decision makers through an IMS.

---

with advanced filtering, browsing and aggregating capabilities.

However, when a large number of ideas are collected, it quickly becomes very time consuming to identify common themes, as well as overlaps, duplicates and related ideas. The system we are developing aims to facilitate this process for the decision maker by providing him or her tools to explore, curate and aggregate ideas. Probabilistic topic models with power-law characteristics are very useful in this context as they enable users and curators to find more relevant and targeted topics, increasing the relevance of retrieved documents and improving their browsing experience.

**About the Author:**

Cédric Archambeau is an area manager at Xerox Research Centre Europe. He also holds an Honorary Senior Research Associate position in the Centre for Computational Statistics and Machine Learning at University College London. His research interests include probabilistic machine learning and data science, with applications in natural language processing, relational learning, personalised content creation and data assimilation.
Medicine Leans on Natural Language Technology to Advance Science and Improve Patient Care

New text mining capabilities open the door to big data analytics.

Health care is a very large domain that covers many disciplines ranging from biomedical research to applied clinical care. While every discipline has its own specific requirements, they all share something in common: a reliance on text for sharing information.

While new forms of communication and information sharing are emerging through the exploitation of electronic databases and IT infrastructures, text and textual ‘documents’ remain the most used and preferred channel for humans to exchange information. In the medical world, this text resides in scientific publications, clinical guidelines, clinical trials, patient records, clinical notes and so on. These documents may be handwritten, typed, semi-structured or completely unstructured.

Although preferred, the use of text has a number of drawbacks. First of all text tends to be ambiguous. When someone says “orange” they may refer to the colour, the fruit, the network provider or (in France) the town. Secondly, when compared to databases where each piece of information is carefully defined, formatted and structured, text is less constrained and “unstructured”. Even after organization into chapters, sections and tables, text remains far less structured than databases. This may not be problematic in hard copy format like a book, but it quickly becomes an issue when we want to search and use the knowledge that lies within such free text. Even when you want to go beyond simple document retrieval using keyword search you need powerful linguistic tools. And dealing with medical textual information goes even beyond being able to process the text. It needs to integrate and be compliant with a wide range of medical standards and terminologies used in the profession such as ICD-10 (International Statistical Classification of Diseases and Related Health Problems) or SNOMED (multilingual clinical health care terminology).

As a result, the development of free text processing technologies has accelerated over the last 10 years giving birth to the creation of robust tools that are now being applied in complex and challenging fields such as medicine.

Identifying when a Paradigm Shift Occurs

In the field of medical research, the ability to read and digest the huge amount of scientific publications that cover recent theories and experiments is a daily challenge. Experts must keep abreast of trends, state of the art methodologies and new discoveries to master
the big picture of current knowledge and thinking. In genomics, progress in DNA sequencing at the beginning of the 2000s created a surge of activity in research and publications related to gene and protein interactions. The number of published results was so large that it was very difficult to capture the exhaustive list needed to build a comprehensive map of interactions. It required advanced text analytics to automatically analyze the content of scientific papers, detect references to gene protein interactions and automatically create such a map.

Similarly, and more recently, new research trends have focused on phenomics which aims at discovering the links between genes and certain types of diseases. Discovering these links requires smart information extraction tools. For example, studies have demonstrated that a single gene is at the origin of certain types of diabetes which sometimes evolve into cancer. Such a link was found thanks to large epidemiological studies which focus on patient records over a long period of time to discover hidden correlations. These studies, which would have required an army of skilled experts to manually annotate relevant pieces of information, can now be conducted by smart information extraction tools to detect, collect and structure this information in an exhaustive and automatic way.

Being immediately informed of new discoveries or major changes that impact knowledge previously taken for granted is of particular importance and very challenging in the medical field due to the amount of information already available. In phenomics the interactions of genes and proteins described in databases are used in the creation of new medication and treatment. It is crucial that these databases be exhaustive, but keeping them up-to-date is extremely labour intensive. Each new publication must be read and compared with existing knowledge to check its relation and possible impact on earlier data. Information extraction technology can be used to detect the paradigm shifts expressed in the language used within the texts. It pinpoints the subtle signs in the syntax or semantics used to describe results that indicate such changes with respect to previous knowledge such as “… a new experiment has demonstrated that molecule X is no longer …” making it easier for the experts to identify new results.

In clinical guidelines, which provide medical best-practices and advice in treating specific diseases, natural language processing can be used to help keep them current. It does so by automatically mining clinical literature, and proposing changes to the experts who manage them. This helps bridge the gap between medical research and medical care. It can also be used to facilitate the enrollment of patients in trials and perform epidemiological studies. In trials, it can analyze and formalize patient eligibility criteria for each trial. It compares this analysis with patient information to identify the individuals who match the criteria or, vice versa i.e., for a given patient it identifies a suitable trial.

**Analysing Risk**

Another important application of text analysis which benefits clinical care activity is risk assessment. Hospitals are very complex environments with multiple practices, expertise, processes and treatments. A shortage of resources, complex processes, emergencies and stress are all factors that, combined, may occasionally lead to human errors that impact patient safety. To prevent or detect these errors as early as possible, a number of monitoring processes are used but the diversity of medical information channels and information systems used in hospitals makes it very difficult. As text is one of the most common channels that medical staff use to describe the complex situations they are faced with, health security experts require tools that can monitor the wide spectrum of data floating around inside hospitals. Hospital acquired infections are a good illustration of such complexity. One of the metrics used to measure the degree of risks of these infections is the amount of soap used per medical staff per surgery. This measure can be cross-checked with actual reports of infections. To better understand the correlation, a French government funded research project was conducted to detect hospital acquired infections from patient records across a group
of hospitals. The objective was to identify from patient discharge summaries any piece of evidence or event that could lead the system to identify where there might be a risk of infection. Researchers worked with the medical staff to translate their medical knowledge into natural language processing rules which could sift through the records and alert staff when an infection was suspected. The system was able to successfully detect 87 percent of infections demonstrating its relevance in accompanying qualified experts to monitor and improve patient care.

**About the Authors:**

Denys Proux is project manager of services and healthcare innovation at Xerox Research Centre Europe. He focuses on applying natural language processing and bioinformatics expertise to create advanced text mining tools for healthcare.

Caroline Hagège is a senior scientist at Xerox Research Centre Europe. She is an expert in parsing in semantics and in the last few years has been researching the applications of this technology combined with medical terminology to analyse electronic medical records and patient data.
Experience Design: The Path from Research to Business

Ever pondered the origin of a seemingly useless feature on your smartphone or wondered how a product manual could be so convoluted? The answer may lie in the gap between research and use of technology in the real world.

With their vision naturally dominated by a technological perspective, researchers and software engineers typically do not make good product designers, and in many cases are not a typical end user of the targeted system.

Experience design bridges the gap by focusing on the quality of the user experience. In this way it can be used to leverage innovation to develop new business opportunities and accelerate the commercialisation of disruptive technology.

The Bridge Between Work Practice and Technology Design

The study of work practices is a method of research that reveals opportunities for innovation - innovation that can be easily adopted by real end users. The studies provide:

- Details on how work is organized. This often includes information about the actual practice itself (legal, medical or other) that would be unknown to computer scientists in distant research laboratories.
- The methods people adopt to achieve what they need to do. Studying these methods reveals a hidden reality: the exceptions, the turnarounds, the barriers, the undocumented work required so that processes run smoothly, at least on the surface level.

Selecting the problems experienced in the workplace and how to address them should be the result of collaboration between interaction and experience designers with computer scientists in multidisciplinary teams. Adding a filter based on “out of the box” thinking typical of designers can help to imagine the future. It can be checked for feasibility by the scientists while still being anchored to real world user problems.

But experience design is not risk free, and must be done carefully to avoid the creation of false expectations. A work practice study can have the drawback that, when exposed as input to computer scientists, it provides a large array of options for improvement. Information and Communication Technology is already far from ideal in the workplace for multiple reasons: organisational practices are not well designed to take into account the features of computer systems, or these systems have often been quite simply poorly designed, and so forth. Many of these problems can be solved through better technology design, whilst others could benefit from the latest technological advances. In our research, we are primarily interested in the latter, as it provides the largest potential and longer term benefit to the user.

Why we Need Designers in Scientific Research.

Beyond user-centred innovation, where design is based on user observation and requirements, there are situations
where researchers need to link technologies to users. This is particularly true with disruptive technologies that can result in new work patterns. Industrial design can help create these links. The challenge is not about making proposals to users, but about imagining new functions or ways they can do their work. This is the sweet spot where industrial design can provide extraordinary insight into the benefits of a disruptive technology, helping users project themselves into a new working environment.

As an example, consider the case of text classification technology. This technology, based on machine learning methods, had been successfully applied for several years as part of document workflow tools, particularly in document imaging and scanning centres. Once scanned document content is analysed and automatically classified for archiving, retrieval etc. Alternative uses were explored by research to see how the technology could be used by knowledge workers in completely different environments. Knowledge workers are subject matter experts who manipulate, analyse and understand information in environments where humans are central to the process. The classification technology can be used to support their work but not to fully automate their processes. When imagining a new way of providing the text classifier to such experts, our main insight was that we needed to deploy the technology in a tangible environment. This was realised through the choice of a multi-touch device to make the interaction fast and intuitive, hiding the technicalities of the underlying statistical algorithms. Turning this vision into reality could only happen by integrating industrial design. A team external to the project, with a fresh eye and different mind-set was required. Industrial designers brought aesthetic and communication skills to the table. Combined with R&D expertise, the team was able to creatively innovate upon the underlying technology realised in a unique jointly patented prototype system. The user interface and system features of this prototype for paralegals in e-discovery can be viewed here.

Using Design to Promote Commercialisation

After the user and the technology comes the third component of experience design - the business. This piece provides the market background research and potentially the offering as well as the development of a final product.

A common challenge for R&D is to convey the value of what is being proposed. The visual part of experience design can help communicate ideas across organisations and businesses, especially when concepts are disruptive and theoretically complex to understand. Being able to walk the business through different scenarios with hands on experimental interaction with the prototype concept helps them to appropriate the technology. It may even generate new ideas for the offer or identify different market opportunities.

There is however an associated risk with this approach. Providing a well-shaped design proposal and working prototype can mislead the business into thinking it is developed and operational which is far from reality. Without a clear communication strategy false expectations may be generated over availability.

Conclusion

By definition research starts from a blank sheet which may be filled with multiple possibilities and many different paths relating to both a technology and business point of view. Experience design can help focus projects on the right sector to pursue based on
user needs or user experience. Furthermore, experience
design provides the tangible and visual support needed
to facilitate the expansion and propagation of the
original idea whilst often generating valuable intellectual
property.

**About the Authors:**

Antonietta Grasso is a senior scientist and Area Manager
at Xerox Research Centre Europe. She has studied for
over twenty years how technology can best support
collaboration and coordination at work.

Yves Hoppenot is in charge of the transfer of
technologies in the domain of managed print services.
He thrives on the everyday challenge aligning research
technology with business needs.

Caroline Privault is a senior project leader and expert in
machine learning at the Xerox Research Centre Europe.
She transfers technologies from research to the business
groups and in particular in the field of eDiscovery for
litigation.
The XRCE timeline tracks and highlights events in the life of the centre. Innovations, collaborations, products, awards and corporate changes.

Visit the online interactive version that includes many more events. As our research focus has evolved we have grouped the last two decades into four broad ‘Themes’ and colour coded the main ‘Research’ competencies within the themes.

**Themes**

- Office of the Future
- Content Management and Printing
- Smarter Document Management℠
- New Services Verticals in Transportation, Healthcare and Customer Care
- Collaboration / Organisational

**Research**

- Natural Language Processing & Information Retrieval
- Computer Vision
- Collaborative and Interactive Systems
- Data Analysis and Process Modelling
1993

Hervé Gallaire appointed Director of new Rank Xerox Research Centre Europe.

1994

Document Satchel prints, transmits and distributes documents.

1996

AT&T translates technical manuals in its automobile industry.

1998

Patrick Bergmans appointed Director of Xerox Research Centre Europe.

2000

Xerox PageCam and Philips Digital Camera launched at CeBit.

2002

Xerox launches MetaWeb.

2004

CategoriX receives Basex Network of Excellence award for technology.

2005

Innovate in France Award.

2006


1993

Bank Xerox Research Centre inaugurated.

1995

RXRC becomes Xerox Research Centre Europe.

1997

RXRC purchased by Xerox.

1999

Xerox introduces coolEye™ 1.0 software.

2001

Xerox awards Digital File Cabinet

2002

1st in TRC benchmark

2004

Innovate in France Award.

2005

1st in EASY evaluation of processing technologies.

2006

Print Infrastructure Mining released in Xerox Office Services 4.2.

Hervé Gallaire appointed Director of new Rank Xerox Research Centre

Document Satchel prints, transmits and distributes documents.

Patrick Bergmans appointed Director of Xerox Research Centre Europe.

Xerox PageCam and Philips Digital Camera launched at CeBit.

Xerox launches MetaWeb.

CategoriX receives Basex Network of Excellence award for technology.

Bank Xerox Research Centre inaugurated.

RXRC becomes Xerox Research Centre Europe.

RXRC purchased by Xerox.

Xerox introduces coolEye™ 1.0 software.

Xerox awards Digital File Cabinet

1st in TRC benchmark

CategoriX receives Basex Network of Excellence award.

Innovate in France Award.

1st in EASY evaluation of processing technologies.

Print Infrastructure Mining released in Xerox Office Services 4.2.


PageCam

Digital Camera

MetaWeb

CategoriX

Innovate in France

EASY evaluation

Print Infrastructure Mining
1991

document management systems

1995

document management systems

1996

Document Management Workshops process over 30 million documents per month.

2006

Xerox Online Support

2007

Xerox launches ‘Virtual Support’ in WorkCentre 7500 series.

2008

Xerox unveils prototype Virtual Help Desk.

This software offers instant diagnostic help to 30 million Xerox multifunction device users around the world.

2009

Xerox launches Open Text
text mining and information
categorization and word spotting

2010

Xerox acquires Affiliated Computer Services (ACS).

Xerox launches $2 million
global research initiative

2011

Xerox online support

2012

DynamicSCT project begins.

Using data analytics and
demand management software,
Xerox acquires Affiliated
Computer Services (ACS).

2013

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2014

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2015

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2016

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2017

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2018

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2019

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2020

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2021

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.

2022

Energy management
software ‘Smoke’ released.

The software uses data
analytics to analyze the use
of printers and determine
how much electricity they
use, allowing organizations
to manage and control their
electricity usage.
2013

**MOBINET EU project begins.**

Building a Europe-wide platform for connected mobility service. 1st XRCE EU project in transportation.

![MOBINET](image)

2013

**Agent Performance Indicator wins Innovation Call Centre Award.**

The software provides all operatives with real-time information on performance, empowering them to adjust their work.

![Agent Performance Indicator](image)

2013

**Xerox in 50 most disruptive companies in MIT Technology Review.**

Innovation in dynamic pricing for parking from Xerox Research Centre Europe disrupting the business.

![Xerox Disruptive Innovation](image)

2013

**XRCE celebrates 20 years of Xerox innovation in Europe!**

![XRCE 20 Years of Innovation](image)
Xerox Research Centre Europe (XRCE)

Overview

Xerox established its European research centre in France in 1993 to create innovative document technology and drive the corporate transition in becoming a services-led technology business.

Today Xerox is the world leader in document management and business process outsourcing. Research in Europe ensures that Xerox maintains that position.

XRCE research covers a broad spectrum of activities linked to information, data, documents and processes. The centre is internationally reputed for its expertise in computer vision, data analytics, natural language processing, machine learning, work practice and process modelling. Our technical expertise is enriched with the knowledge gathered from our sociologists and ethnographers who develop a deep understanding of how our customers work to combine both human and machine elements.

Scientists and their research are consistently recognized and awarded in international competitions and forums. Technologies developed in the centre are part of the Xerox services offerings and strategically contribute to the competitiveness of Xerox customers.

Alongside its research and development the centre runs a global Customer Led Innovation program to co-create services with customers. It also hosts the European Technology Showroom, a showcase for Xerox research and an exchange forum for scientists, engineers and customers.
Combining scientific vision with business reality.
Natural Language Processing & Information Retrieval

Overview

Text analysis is at the root of the tools that help us extract, analyse, understand, generate or translate information. Research in natural language processing was one of the first research areas to be established at the centre with a focus on multilingual documents and document processes which form the backbone of many global enterprise operations. This work led to the creation of basic tools for linguistic analysis such as tokenizers, morphological analysers and parsers. These components are used to develop descriptions of various languages and the relation between them. They are then integrated into higher level applications such as terminology extraction, information retrieval or translation aid. Tools for 20 languages were developed, including non-European languages such as Arabic and Chinese. Projects included document authoring and translation support first used in 1996 within the authoring and translation environment ATOLL and commercialised 2 years later as the Xerox Translation & Authoring Systems ‘XTRAS’ using the XeLDA linguistic platform.

In the late 1990s there was also a good deal of seminal research on recommender systems, which subsequently became famous in the form of collaborative filtering on sites like amazon.com. Such interest in web applications culminated in our researchers introducing some of the world’s first tools for metasearch or federated search engines, leading to the foundation of the Xerox company AskOnce in 1999.

In the last decade our research in natural language has advanced using two complementary approaches; machine learning methods and rule based parsing and semantics. Machine learning research creates models and algorithms for multilingual and multimodal document categorization and clustering, social network analysis, novelty detection and advanced statistical machine translation. The rule-based Xerox Incremental Parser is a robust parsing formalism and engine that can describe fine-grained linguistic phenomena through syntactic dependencies. It is integrated with a number of advanced components for semantic disambiguation, co-reference resolution, discourse analysis and temporal processing to be able to take into account the semantic context of language. The content discovery software named ‘FactSpotter’ was developed using these tools to discover linguistic relationships and meaning across document collections. It has been successfully piloted in hospitals to analyse patient records and help detect hospital acquired infections. Other successful applications in natural language that are part of the Xerox Smarter Document Management tools range from enterprise mailroom automation, where more than 30 million documents are automatically processed every month, to our award-winning document categorization used in eDiscovery for corporate litigation processes. Since the acquisition of Affiliated Computing Services in 2010, the centre’s natural language technology is also being applied to other areas of healthcare and in query automation and sentiment analysis in customer care. Current research is integrating our fine-grained linguistic and semantic models into

---

1 See the 20th anniversary articles in ‘Computer Science Meets Journalism’, ‘Can ‘made up’ artificial languages help computers translate real ones?’, ‘Will translation ever be fully automated?’, ‘Do Computers Need Grammar Books Too?’ and ‘How Natural Language Technology is Helping Advance Medical Science and Patient Care’. 
machine learning methods to improve the state-of-the-art in multilingual language understanding and analysis.

2004

**CategoriX receives Basex Excellence award for technology.**

Industry analyst firm selects Xerox for its invention of an automated information categorizer. The categorizer can “read” an electronic document, determine classification by subject, and then route it to the appropriate individual via e-mail - all without human intervention.

2011

**FactSpotter selected by Premier Healthcare alliance as one of the latest advancements revolutionizing healthcare.**

Premier Innovation Celebration Nashville June 2011.
Natural Language Processing & Information Retrieval
A Selection of Patents

7672830
Apparatus and Methods for Aligning Words in Bilingual Sentences
Cyril Goutte, Michel Simard, Kenji Yamada, Eric Gaussier, Arne Mauser

7630977
Categorization Including Dependencies Between Different Category Systems
Eric Gaussier, Jean-Michel Renders, Cyril Goutte, Caroline Privault

7536295
Machine Translation Using Non-Contiguous Fragments of Text
Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Michel Simard, Kenji Yamada

7720848
Hierarchical Clustering with Real-Time Updating
Agnes Guerraz, Caroline Privault, Cyril Goutte, Eric Gaussier, François Pacull, Jean-Michel Renders

8239379
Semi-Supervised Visual Clustering
Boris Chidlovskii, Loïc Lecerf

8374844
Hybrid System for Named Entity Resolution
Caroline Brun, Maud Ehrmann, Guillaume Jacquet

8103669
System and Method for Semi-Automatic Creation and Maintenance of Query Expansion Rules
Stefania Castellani, Aaron Kaplan, Frédéric Roulland, Jutta K Willamowski

8189930
Categorizer with User-Controllable Calibration
Jean-Michel Renders, Caroline Privault, Eric Cheminot

8209599
Method and System for Handling References in Markup Language Documents
Jean-Yves Vion-Dury, Jean-Pierre Chanod

8165974
System and Method for Assisted Document Review
Caroline Privault, Jacki O’Neill, Jean-Michel Renders, Victor Ciriza, Yves Hoppenot

6681369
System for Providing Document Change Information for a Community of Users
Jean-Luc Meunier, Damian Arregui, Natalie S Glance

6901360
System and Method for Transferring Packed Linguistic Structures
Marc Dymetman, Frédéric Tendeau

7933454
Class-Based Image Enhancement System
Marco Bressan, Christopher R Dance, Gabriela Csurka

7620539
Method and Apparatus for Identifying Bilingual Lexicons in Comparable Corpora
Eric Gaussier, Jean-Michel Renders, Hervé Déjean, Cyril Goutte, Irina Matveeva

6381598
System for Providing Cross-Lingual Information Retrieval
Jutta K Willamowski, Uwe M Borghoff
7058567
Natural Language Parser
Salah Aït-Mokhtar, Jean-Pierre Chanod, Claude Roux

8108766
Xpath-Based Display of a Paginated XML Document
Jean-Luc Meunier

8285541
System and Method for Handling Multiple Languages in Text
Caroline Brun

6473729
Word Phrase Translation Using a Phrase Index
Michel Gastaldo, Gregory T Grefenstette

6405162
Type-Based Selection of Rules for Semantically Disambiguating Words
Frédérique Segond, Caroline Brun

7296223
System and Method for Structured Document Authoring
Boris Chidlovskii, Hervé Déjean

6829599
Method and System for Improving Answer Relevance in Meta-Search Engines
Boris Chidlovskii

8009921
Context Dependent Intelligent Thumbnail Images
Gabriela Csurka

7035841
Method for Automatic Wrapper Repair
Boris Chidlovskii

7139754
Method for Multi-class, Multi-Label Categorization Using Probabilistic Hierarchical Modeling
Cyril Goutte, Eric Gaussier

7165216
Systems and Methods for Converting Legacy and Proprietary Documents into Extended Mark-Up Language Format
Boris Chidlovskii, Hervé Déjean

6289304
Text Summarization Using Part-of-Speech
Gregory T Grefenstette
Natural Language Processing & Information Retrieval

Publications

20 selected publications from 1993 - 2013

**Citizen Opinion Summarization for Electronic Governance by Harnessing the Crowd**
To appear in ACM Transactions on Knowledge Discovery from Data.
Darling W., Bouchard G.

**Learning Opinionated Patterns for Contextual Opinion Detection**
Brun C.

**Prediction of Learning Curves in Machine Translation**
Kolachina P., Cancleda N., Dymetman M., Venkatapathy, S.

**Full and Mini-Batch Clustering of News Articles with Star-EM**
European Conference on Information Retrieval (ECIR), Barcelona, Spain. April 1-5, 2012.
Gallé M., Renders J.M.

**Pushing the Frontier of Statistical Machine Translation: Preface**
Specia L., Cancleda N.

**Learning Recommendations in Social Media Systems by Weighting Multiple Relations**
Chdlovskii B.

**Information-Based Models for Ad Hoc IR**
Clinchant S., Gaussier E.

**Learning Machine Translation**
Goutte C., Cancleda N., Dymetman M., Foster G. (editors)

**Estimating the Sentence-Level Quality of Machine Translation Systems.**
Specia L, Turchi M., Cancleda N., Dymetman M., Cristianini N.

**A Hybrid System for Named Entity Metonymy Resolution**
Brun C., Ehrmann M., Jacquet G.
XIP Temporal Module: TempEval Campaign
Hagege C., Tannier X.

Discourse and Citation Analysis with Concept-matching
Sandor A., Kaplan A., Rondeau G.

Translating with Non-contiguous Phrases
Simard M., Cancedda N., Cavestro B., Dymetman M., Gaussier E., Goutte C., Langlais P., Yamada K., Mauser A.

Word-Sequence Kernels
Cancedda N., Gaussier E., Goutte C., Renders J.M.

Robustness beyond Shallowness: Incremental Deep Parsing
Journal of Natural Language Engineering 8:3 pp. 121-144 (June 2002).
Aït-Mokhtar S., Chanod J.P., Roux C.

A Hierarchical Model for Clustering and Categorising Documents
Gaussier E., Goutte C., Popat K., Chen F.

Semantic Encoding of Electronic Documents
Brun C., Segond F.

Incremental Finite-state Parsing
Aït-Mokhtar S., Chanod J.P.
Computer Vision
Overview

Computer vision research is concerned with understanding the content of visual data such as photos, videos or scanned documents. In the early 90’s XRCE research in this field focussed on novel ways of working with paper and the interconnectivity of multimedia to create media spaces. The ‘digital desk’ used a video camera and projector over a desk to provide ‘real’ desktop interaction, including both interaction with physical items on the desktop as well as virtual projected ones. The inadequate resolution of cameras at the time involved considerable innovation in image processing techniques to overcome lighting and resolution difficulties and to restore images that were blurred and patchy. These were used in the Camworks prototype which performed face up document scanning and was commercialised in collaboration with Philips as the PageCam product in the year 2000. These new techniques were also used to improve the quality of documents scanned with multifunction devices to enable the automation of many document processes.

This vision of giving the gift of sight to computers continues today. Combined with textual information from the images themselves (e.g. OCR, natural text) or external sources (web pages, internal database records) our research deals with all types of interaction where visual content is one component. The main areas of focus are robust image signatures, large scale classification and retrieval, image segmentation and detection and leveraging multi-modal Information. We were among the first groups in the world to start combining machine learning with computer vision and were the first to propose the “bag-of-visual words model for visual categorization” which has since become a de facto standard in the community.

Our research feeds into current business and blue-sky future disruptive technologies including Smarter Document ManagementSM, Healthcare applications and Transportation Solutions.

---

1 See anniversary article ‘A tribute to visual words and how they revolutionized computer vision’
2011

Image categorisation technology wins Imagenet benchmarking challenge.

The benchmark is part of the PASCAL network challenge for visual objects.

2006


AIE uses algorithms developed by XRCE to automatically improve images before printing.
Computer Vision
A Selection of Patents

7907786  Red-Eye Detection and Correction
Lixin Fan, Jutta K. Willamowski, Christopher R Dance

7426312  Contrast Enhancement of Images
Christopher R Dance, Damian Arregui

7826665  Personal Information Retrieval Using Knowledge Bases for Optical Character Recognition Correction
Marco Bressan, Hervé Dejean, Christopher R Dance

7756341  Generic Visual Categorization Method and System
Florent Perronnin

7639893  Histogram Adjustment for High Dynamic Range Image Mapping
Jiang Duan, Marco Bressan, Christopher R Dance, Guoping Qui

7680341  Generic Visual Classification With Gradient Components-Based Dimensionality Enhancement
Florent Perronnin

8111923  System and Method for Object Class Localization and Semantic Class Based Image Segmentation
Gabriela Csurka Florent Perronnin

8379974  Convex Clustering for Chromatic Content Modeling
Naila Murray, Florent Perronnin, Luca Marchesotti, Sandra Skaff

8532399  Large Scale Image Classification
Florent Perronnin, Jorge Sánchez, Thomas Mensink

5748805  Method and Apparatus for Supplementing Significant Portions of a Document Selected without Document Image Decoding with Retrieved Information
M Margaret Withgott, William M Newman, Steven C Bagley, Daniel P Huttenlocher, Ronald M Kaplan, Todd A Cass, Per-Kristian G Halvorsen, John S Brown, Martin Kay

6493469  Dual Video Camera System for Scanning Hardcopy Documents
Michael J Taylor, William M Newman

6577762  Background Surface Thresholding
Mauritius Seeger, Christopher R Dance

6463220  Method and Apparatus for Indicating a Field of View for a Document Camera
Christopher R Dance, Mauritius Seeger

7123292  Mosaicing Images with an Offset Lens
Mauritius Seeger, Christopher R Dance, Stuart A Taylor, William M Newman, Michael J Flynn
6766069
Ext Selection from Images of Documents Using Auto-Completion
Christopher R Dance, William M Newman, Alex S Taylor, Stuart A Taylor

7505607
Identifying Objects Tracked in Images Using Active Device
Jean-Luc Meunier, Frédéric Roulland, Alba Ferrer-Biosca, James L Crowley

7574349
Statistical Language-Model Based System for Detection of Missing Attachments
Florent Perronnin

6067112
Interactive Desktop Display System for Automatically Adjusting Pan and Zoom Functions in Response to User Adjustment of a Feedback Image
Pierre D Wellner, Michael J Flynn, Kathleen A Carter, William M Newman

6640010
Word-To-Word Selection on Images
Mauritius Seeger, Christopher R Dance, Stuart A Taylor, William M Newman

8533204
Text-Based Searching of Image Data
José-Antonio Rodriguez Serrano, Florent Perronnin, Craig J Saunders

6922487
Method and Apparatus for Capturing Text Images
Christopher R Dance, Guy Ruddock, William M Newman, Stuart A Taylor
Computer Vision

Publications

20 selected publications from 1993 - 2013

Image Classification with the Fisher Vector: Theory and Practice
Sanchez J., Perronnin F., Mensink T., Verbeek J.

What is a Good Evaluation Measure for Semantic Image Segmentation
Csurka G., Larlus D., Perronnin F.

A Model-based Sequence Similarity with Application to Handwritten Word-spotting
Rodriguez J.A., Perronnin F.

Leveraging Category-Level Labels for Instance-Level Image Retrieval
Gordo A., Rodriguez J.A., Perronnin F., Valveny E.

Data-driven Vehicle Identification by Image Matching
Rodriguez J.A., Sandhawalia H., Bala R., Perronnin F., Saunders C.

Document Classification Using Multiple Views
Best paper ‘Nakano award’.
Gordo A., Perronnin F., Valveny E.

Assessing the Aesthetic Quality of Photographs using Generic Image Descriptors
Marchesotti L., Perronnin L., Larlus D., Csurka G.

Towards Automatic Concept Transfer
Non-Photorealistic Animation and Rendering (NPAR), Vancouver, Canada. August 5-7, 2011.
Murray N., Skaff S., Marchesotti L., Perronnin F.

Semantic Combination of Textual and Visual Information Multimedia Retrieval
Ah-Pine J., Clinchant S., Csurka G.

A Framework for Visual Saliency Detection with Applications to Image Thumnailing
Marchesotti L., Cifarelli C., Csurka G.
Fisher Kernel on Visual Vocabularies for Image Categorization
Csurka G., Perronnin F.

Adapted Vocabularies for Generic Visual Categorization
European Conference on Computer Vision (ECCV), Graz, Austria. May 7-13, 2006.
Perronnin F., Dance C.R., Csurka G., Bressan M.

Visual Categorization with Bags of Keypoints
Dance C.R., Willamowski J., Fan L., Bray C., Csurka G.

Perspective Estimation for Document Images
Dance, C.R.

Binarising Camera Images for OCR
Seeger M., Dance C.R.

Estimation of The Parameters of Skewed Alpha-stable Distributions
Dance, C. R., Kuruoglu E
Collaborative and Interactive Systems

Overview

Research in collaborative and interactive systems spans a fairly broad sector of research conducted at the centre over the past twenty years. A common thread that runs through most of the projects in this area is that of the human element to ensure that what is developed is both useful and usable by end users.

Researchers realised when mobile devices first appeared that they provided a huge opportunity yet a real technical challenge to support the rapidly growing number of mobile workers. Some of the distinctive challenges included the unpredictability of document needs, difficulties accessing remote documents and secure access from unfamiliar locations. The document ‘Satchel’ was one application developed to meet these challenges in 1994. It was commercialised as MobileDoc to e-mail and fax documents remotely from cell phones, handheld computers and pagers. Mobile printing was later added to these capabilities and is a core part of the Xerox offering today.

In parallel, early studies in the Office of the Future showed that, despite the growth in digital documents, paper would have a fundamental role for the foreseeable future although the way it would be used would radically change. The research challenge was therefore not how to replace paper, but how to support the smooth flow of information across the paper-digital boundary. We began with looking at ways of integrating paper and electronic environments and how technology could support novel ways of working with paper. Numerous technologies and applications resulted from this work ranging from multimedia interactive systems such as the ‘DigitalDesk’, to encoding and automatically extracting information from paper to access, repurpose and above all, share its content (e.g. SmartPrinter, CopyFinder).

This theme of knowledge sharing was pursued with the explosion of the web and the birth of social and virtual expert communities by building recommender systems such as the ‘Knowledge Pump’ in the late 90s or today’s collaborative idea management systems for governments that collect, analyse and interpret opinions expressed on the Internet.

Much of the research conducted in this area is based on the fieldwork and ethnographic studies of work sites and organisations. By detailed observation and participation in the customer sites, very detailed descriptions of customer work and organisations are built up “from the inside”. Results of these studies are used to support decisions about future services, product requirements and directions and to identify the potential for more speculative new technologies. Important in depth fieldwork has been carried out in document intensive environments such as the IMF and Department of Work and Pensions in the UK, print shops and graphic design agencies, home offices, paralegal work and customer care centres. The remote call assistance project investigated how new and improved ‘instruments’ for self-troubleshooting and remote technical support could be developed to better assist in office device repair. A number of new technological solutions based on our research were produced including three award winning systems: the Xerox online support system that can be
accessed directly from WorkCenter 7500 printer screens, the Virtual Help Desk troubleshooting system that offers help via 3D diagnostics and instant live support directly from the printer and the Agent Performance Indicator technology which provides personal real time performance status to call centre staff on a mini dashboard¹. Research is currently being pursued in collective intelligence and crowdsourcing and how they can be used to improve business processes.

¹ See anniversary articles ‘Do you act like your computer (designer) thinks you ought to?’ and ‘Experience design: the path from research to business.’

2010

Xerox online support solution wins TSIA Innovation Award.

The Technology Services Industry Association recognized the software which allows customers to interact with the Xerox knowledge base directly on the printer screen.

2011

Wall Street Journal 2011 Innovation Award.

Virtual Help Desk technology from XRCE integrates 3D and VoIP to connect users to call centre agents directly from the multifunction device.
Collaborative and Interactive Systems
A Selection of Patents

7562223
Multifunction Device with Secure Job Release
François Ragnet, Victor Ciriza, Olivier Fambon, Yves Hoppenot

7822597
Bi-Dimensional Rewriting Rules for Natural Language Processing
Caroline Brun, Caroline Hagège, Claude Roux

7664629
Second Language Writing Advisor
Marc Dymetman, Pierre Isabelle

8280842
Collaborative Linking of Support Knowledge Bases with Visualization of Device
Frédéric Roulland, Pascal Valobra, Ye Deng, Stefania Castellani, Jacki O’Neill

8482556
Detachable Screen for Multifunction Device Showing 3D Dynamic Views
Jean-Luc Meunier, Stefania Castellani, Victor Ciriza, Denys Proux

7495792
Programmable Physical Document
David Snowdon, Christer Fernström, Marc Dymetman, Natalie S Glance

8384941
System and Method for Enabling an Environmentally Informed Printer Choice at Job Submission Time
Victor Ciriza, Maria Antonietta Grasso, Yves Hoppenot, Jutta K. Willamowski

7346841
Method and Apparatus for Collaborative Annotation of a Document
Ercan E Kuruoglu, Alex S Taylor, Mauritius Seeger, Stuart A Taylor

6868424
Electronic Filing System with File-Placeholders
Rachel Jones, Allan Maclean, Richard Bentley, Kevin Palfreyman, James Pycock, Jon O’Brien, Graham Button

409452
Method and Apparatus for Controlling Document Service Requests from a Mobile Device
François Ragnet, Christer Fernström

6005547
Calibration of an Interactive Desktop System
William M Newman, Quentin Stafford-Fraser, Richard Bentley
6515988
Token-based Document Transactions
Margery A Eldridge, Michael J Flynn, Christopher M Jones, Michiel Kleyn, Michael G Lamming, David L Pendlebury

6473523
Portable Text Capturing Method and Device Therefor
William M Newman, Carl Bothner, Ben S Wittner

7827026
Bilingual Authoring Assistant for the ‘Tip Of The Tongue’ Problem
Caroline Brun, Marc Dymetman, Frédérique Segond

8351075
Print Mediator

8022823
Serendipitous Repair of Shared Device
Jean-Luc Meunier

8390886
System and Method for Print Profile Selection
Jutta K. Willamowski Frédéric Roulland, David Martin

6493760
Standalone Device for Identifying Available Document Services in a Token Enabled Operating Environment
David L Pendlebury, Michael G Lamming, Margery A Eldridge, Christopher M Jones, Mark Stringer

6922725
Method and Apparatus for Processing Document Service Requests Originating From a Mobile Computing Device
Michael G Lamming, Allan Maclean, Anthony F Frayling

6947609
System with Motion Triggered Processing
Mauritius Seeger, Stuart A Taylor, Christopher R Dance

6622123
An Interactive Translation System and Method
Jean-Pierre Chanod, Marc Dymetman

6735622
Transferring Constraint Descriptors for Documents
Jean-Marc Andreoli, Uwe M Borghoff

6345304
Obtaining Network Addresses From Identifiers
Marc Dymetman, Max Copperman

6421716
System for Generating Context-sensitive Hierarchically Ordered Document Service Menus
Margery A Eldridge, Michael J Flynn, Christopher M Jones, Michiel Kleyn, Michael G Lamming, David L Pendlebury

8380743
System and Method for Supporting Targeted Sharing and Early Curation of Information
Gregorio Convertino, Ed H Chi, Benjamin V Hanrahan, Nicholas Kong, Guillaume Bouchard, Cédric Archambeau

6973655
System and Method of Integrating Software Components
Thierry Jacquin, Michel Gastaldo
Collaborative and Interactive Systems

Publications

20 selected publications from 1993 - 2013

**Relationship-based Business Process Crowdsourcing?**
IFIP Conference on Human-Computer Interaction, Cape Town, South Africa, September 2 - 6, 2013.
O’Neill J., Martin D.

**Lifting the Mantle of Protection from Weber’s Presuppositions in his Theory of Bureaucracy.**
Button G, Martin D., O’Neill J., Colombino T.

**Collective Intelligence in Organizations: Tools and Studies Introduction**
Grasso A., Convertino G.

**A Little Knowledge is a Dangerous Thing**
O’Neill J., Martin M., Colombino T., Grasso, A.

**From Ethnographic Study to Mixed Reality: A Remote Collaborative Troubleshooting System**
SIGCHI “Best of CSCW” award.
O’Neill J., Castellani S., Roulland F., Juliano C., Dai L., Hairon N.

**A Semantic System to Support Legal Case Construction**
Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) Conference Revised Selected Papers, Communications in Computer and Information Science (CCIS Series), Springer, Vol. 348, pp. 217-231.
Castellani S., Lagos N., Grasso A., Hairon N., Martin D., Segond F.

**Talking about (my) Generation: Creativity, Practice, Technology & Talk**
Martin D., O’Neill J., Randall D.

**Designing Technology as an Embedded Resource for Troubleshooting**
Castellani S., Grasso A., O’Neill J., Roulland F.

**Creation and Maintenance of Query Expansion Rules**
International Conference on Enterprise Information Systems (ICEIS), Milan, Italy. May 6-10, 2009.
Best paper award in the area of human-computer interaction.
Castellani S., Kaplan A., Roulland F., Willamowski J., Grasso A.

**Colour Management is a Socio-technical Problem**
Computer Supported Cooperative Work (CSCW), San Diego, California, USA. Nov 8-12, 2008.
O’Neill J., Martin D., Colombino T., Roulland F., Willamowski J.
Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation
IEEE Transactions on Knowledge and Data Engineering (KDE), Vol: 19, Issue: 3 pp. 355-369. 2007.
Pirotte A., Renders J.M., Saerens M.

Unremarkable Computing
Tolmie P., Pycock J., Diggins T., MacLean A., Karsenty A.

Pollen: Using people as a communication medium
Glance N., Snowdon D., Meunier J.L.

Dourish P., Button G.

Intelligent Paper
Dymetman M., Copperman M.

A Comparison of Reading Paper and On-Line Documents
O’Hara K., Sellen A.

Inside the IMF: An Ethnography of Documents, Technology and Organisational Action.
Harper, R.

Interactive System Design
Newman, W. & Lamming, M.

Interacting with Paper on the DigitalDesk
Communications of the ACM, Volume 36 (7), 1993, pp. 87-96. Special issue on computer augmented environments: back to the real world.
Wellner P.

Bringing Media Spaces into the Real World
Pagani D., Mackay, W.
Data Analysis and Process Modelling

Overview

The range of XRCE’s work on data analysis and process modelling is extremely broad and our activities in this area continue to grow in line with the increasing popularity of the term “analytics.”

To trace the evolution of our work in this field it began with a decade of research on document layout understanding, where we took the novel twist of trying to infer the structure of documents from PDF files, enabling us to achieve more accurate inferences than OCR companies who had always worked with scanned documents. This work found numerous applications with our document services outsourcing business, and we have made some of it freely available in the application pdf2xml at sourceforge.net1.

Similarly, we applied our machine learning and dynamic optimization skills to numerous problems for the managed print services business leading to device-usage models, failure detection methods, stocking and ordering methods for supply chains and intelligent time-out mechanisms.

As we became increasingly a “services company” and expanded our activities to many more verticals (e.g. public transport, parking, call centres) in 2010, we have been confronted with an ever broader range of problems. Nevertheless these problems frequently share common elements.

One such common theme is Bayesian inference, which is all about optimal reasoning under uncertainty given data and prior knowledge. Our approaches frequently focus on making such problems computationally tractable, for instance using variational methods which approximate the optimal beliefs by much simpler classes of probability distributions.

Another such theme is mechanism design which is about “inverse game theory” or “how to design a game so that agents’ optimal strategies have certain properties” and which is an appropriate foundation for studying pricing (e.g. for parking2) and incentive mechanisms (e.g. for call centre agents).

The final such theme stems from the fact that services companies rely strongly on processes and therefore, they permanently face a trade-off between: reducing costs and simplifying peoples’ lives by standardising and automating those processes; specialising processes to enable optimal performance for a particular customer’s needs; and retaining the agility to rapidly respond to new customers, new problems and new software components. In response, we have designed methods for maintaining and managing schemas, domain specific languages and process modelling environments.

The worldwide volume of data and the pervasiveness of integrated and automated processes have grown tremendously over the last twenty years and the end to that growth is nowhere in sight, which means that XRCE’s research on data analysis and process modelling will continue to intensify for the foreseeable future.

1 See anniversary article “Monks and Markup”.
2 See anniversary article “How economics and machine learning can tackle transportation congestion.”
2006
Print Infrastructure Mining released in Xerox Office Services 4.2.

Data analytics technology supports intelligent decisions to optimize a customer’s printing environment to maximize cost savings and end-user productivity.

2013
Xerox in 50 most disruptive companies in MIT Technology Review.

Innovation in dynamic pricing for parking from Xerox Research Centre Europe disrupting the business.
Data Analysis and Process Modelling

A Selection of Patents

7937653
Method and Apparatus for Detecting Pagination Constructs Including a Header and a Footer in Legacy Documents
Hervé Déjean, Jean-Luc Meunier

7715037
Bi-Directional Remote Visualization for Supporting Collaborative Machine Troubleshooting
Stefania Castellani, Victor Ciriza, Maria Antonietta Grasso, Jacki O’Neill, Peter Tolmie, François Ragnet, Frédéric Roulland

7567946
Probabilistic Modeling of Shared Device Usage
Jean-Marc Andreoli, Guillaume Bouchard

7865089
Soft Failure Detection in a Network of Devices
Jean-Marc Andreoli, Guillaume Bouchard, Victor Ciriza

7929165
Method and System for Controlling Printer Utilization in a Networked Environment
Marco Bressan, Guillaume Bouchard

7707493
Method for Generating Presentation Oriented XML Schemas Through a Graphical User Interface
Jean-Yves Vion-Dury, Jean-Pierre Chanod

8074124
Visualization of User Interactions in a System of Networked Devices
Guillaume Bouchard, Laurent Donini, Pascal Valobra, Victor Ciriza

8230248
Printer Time-out
Christopher R Dance, Victor Ciriza, Laurent Donini

8339680
Printer Image Log System for Document Gathering and Retention
Caroline Privault, Victor Ciriza, Reave Reisler, Jean Ellefson

8159493
Adaptive Grand Tour
Loïc Lecerf, Guillaume Bouchard

8429001
Limited Lottery Insurance
Christopher R Dance, Onno Zoeter, Guillaume Bouchard

8260655
Price Optimization with Robust Learning Constraint
Christopher R Dance, Onno Zoeter

8442861
Multi-Dimensional Price Determination
Christopher R Dance, Onno Zoeter

7454400
System for Negotiation With Mirroring
Jean-Marc Andreoli, Stefania Castellani

7577706
Workflow System and Method
Damian Arregui, François Pacull, Jutta K. Willamowski, Stefania Castellani

5539665
Recording and Retrieval of Information Relevant to the Activities of a User
Michael G Lamming, Michael J Flynn

7849398
Decision Criteria for Automated Form Population
Sébastien Dabet, Marco Bressan, Hervé Poirier
6732361
Generating Combinations of Offers and Using Action Identifiers From the Offers to Obtain Performance of Combinations of Actions
Jean-Marc Andreoli, Francois Pacull, Jean-Luc Meunier

7129891
Method for Determining Proximity of Devices in a Wireless Network
Jean-Luc Meunier

7536369
XML-Based Architecture for Rule Induction System
Hervé Déjean

8145992
Validation Assisted Document Conversion Design
Thierry Jacquin, Jean-Pierre Chanod
20 selected publications from 1993 - 2013

A Platform-Independent Mechanism for Deployment of Business Processes Using Abstract Services
Mos A., Jacquin, T.

A Generic Calculus of XML Editing Deltas
Vion-Dury J.Y.

Secured Management of Online XML Document Services through Structure Preserving Asymmetric Encryption
Archambeau C., Guo S., Zoeter O.

Sparse Bayesian Multi-Task Learning
Archambeau C., Guo S., Zoeter O.

Optimal and Robust Price Experimentation: Learning by Lottery
Dance C.R., Zoeter, O.

Robust Bayesian Matrix Factorisation
Lakshminarayanan B., Bouchard G., Archambeau C.

Numbered Sequence Detection in Documents
Déjean H.

Xeproc: A Model-Based Approach towards Document Process Preservation
Déjean H., Jacquin T., Chanod J.P.

Automated Quality Assurance for Document Logical Analysis
Meunier J.L.

Variational Bounds for Mixed-Data Factor Analysis
Khan E., Bouchard G., Marlin B., Murphy K.
Multi-modality in One-class Classification
Hovelynck M., Chidlovskii B.

On Tables of Contents and How to Recognize Them
Déjean H., Meunier J.L.

Managing XML References through the XRM Vocabulary
Balisage: The Markup Conference, Montréal, Canada, August 11-14, 2009.
Vion-Dury J.Y.

Split Variational Inference
Bouchard G., Zoeter, O.

London: Routledge. 1993
Button, G. (Ed.)

A System for Converting PDF Documents into Structured XML Format
Déjean H., Meunier J.L.

Optimized XY-cut for Determining a Page Reading Order
Meunier J.L.

From Legacy Document to XML: A Conversion Framework
Chanod J.P., Chidlovskii B., Dejean H., Fambon O., Fuselier J., Jacquin T., Meunier J.L.

Logic-based XPath Optimization
Genevès P., Vion-Dury J.Y.

Focussing and Proof Construction
Andreoli J.M.
It is often said that in research, the quality and expertise of your people is even more important than in most other disciplines. At XRCE our ethnographers and social scientists would argue that people are your major asset regardless of your business and they would be right.

What we focus on is getting the right balance - the right balance of skills, across and within disciplines.

The right balance of diversity in all its shapes and forms. We do our best to provide a working environment that embraces this diversity, which we believe gives rise to outstanding results. And that’s precisely what our people enjoy most; working on challenging topics that have real world impact.

The faces you see here are those of many of our researchers, engineers, doctoral students and support staff. Each and every one is an XRCE ambassador.

Our goal is to make Xerox a great place to work. Through a comprehensive set of employee-focused initiatives, we promote diversity by nurturing a culture of inclusion and opportunity, and through measurable actions.

Xerox complies with Equal Employment Opportunity (EEO) guidelines. Xerox does not discriminate on the basis of race, colour, religious belief, sex, age, national origin, citizenship status, marital status, union status, sexual orientation, gender identity or individuals with a disability.

Our main building is compliant with the AGEFIPH standards for disabled access.
A good day in the Technology Showroom is a day spent with a customer. On average 800 customers participate in workshops organized by the Showroom team every year either at XRCE or across Europe which fortunately means there are many such days. The Showroom allows Xerox research laboratories from around the world to demonstrate their innovative ideas in a professional, “hands-on” environment. The team also facilitates “Dreaming Sessions” where customers and prospects are invited to share their top challenges with Xerox researchers working on the cutting edge of technology.

The Technology Showroom has a dual mission: to showcase leading-edge technologies from Xerox labs to collect usage scenarios and gather insight for further research and development from potential users, and secondly to accelerate the sales cycle by strengthening the relationships with C-Level clients via customized client experiences.

For almost 20 years the XRCE Technology Showroom has accompanied Xerox in its transformation into a services-led company and has adapted its activities to market pressures and changes Xerox has undergone over the last two decades.

New technologies are best understood when encountered in a setting which relates to everyday life and work. The Technology Showroom has always helped Xerox clients, corporate officers, business leaders, analysts and the sales force understand what the future would be like if the visions’ of our researchers were made real. This has driven us to build and demonstrate a wide range of topics including the Digital Workplace, Smart Cities and Business Process Transformation. The Technology Showroom team uses all forms of media communications to convey XRCE’s vision of the future.

Customer-led innovation will be at the heart of the next disruptive business, and the XRCE Technology Showroom will contribute by creating immersive environments to showcase innovation and help identify unmet client needs.

Upcoming demonstrations and events will continue to be one of the most effective ways of generating immediate “Voice of the Customer” feedback on applications of emerging technologies. The XRCE Technology Showroom is a must for every customer willing to partner with Xerox to develop new solutions and services based on leading edge innovation.
Open and Collaborative Innovation

While drawing on the strength of the Xerox Corporation around the world, our focus is on Europe. We work with organisations drawn from the continent and reflect their needs and talents. We cooperate with the scientific community and collaborate with a wide range of European research organisations particularly within the European Community frameworks but also as part of national government initiatives.
EU Research Projects

Framework Programme 7 (2007-2013)

PERICLES
Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics. Ensuring digital assets (e.g. data sets, software, media, etc.) created today are available and useful for future generations. 2013-2017. (Information and Communication Technologies-Collaborative Project, n°601138)

MOBINET
Europe-Wide Platform for Cooperative Mobility Services. Develop, deploy and operate the technical and organisational foundations of an open, multi-vendor platform for Europe-wide mobility services. 2012-2016. (Information and Communication Technologies - Collaborative Project, n°318485).

FUSEPOOL
Fusing and Pooling information for product/service development and research. Defines and enriches raw data using common standards and provides tools to analyse and visualize data so that end users and other software receive timely, context-aware and relevant information. 2012-2014. (Information and Communication Technologies - Collaborative Project, n°296192)

EURECA
Enabling information re-Use by linking clinical Research and Care. An advanced, standard-based and scalable semantic integration environment to bridge the gap between clinical research and clinical care systems. 2012-2015. (Information and Communication Technologies - Collaborative Project, n°288048)

TRANSLECTURES
Transcription and Translation of Video Lectures. Developing innovative, cost-effective solutions to produce accurate transcriptions and translations in VideoLectures.NET, with generality across other Matterhorn-related repositories. 2011-2014. (Information and Communication Technologies - Collaborative Project, n°287755)

FUPOL
Future Policy Modelling a completely new approach to traditional politics. Major innovations like multichannel social computing and crowd sourcing will change the way politicians communicate with citizens and enterprises and take decisions. 2011-2015. (Information and Communication Technologies - Collaborative Project, n°287119)

ORGANIC LINGUA
Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education. An automated multi-lingual service that aims to facilitate the usage, exploitation and extension of digital educational content related to Organic Agriculture and Agroecology. 2011-2014. (Competitiveness Innovation Framework Program-ICT Policy Support Program, n°270999)

GALATEAS
Generalized Analysis of Logs for Automatic Translation and Episodic Analysis of Searches analyses user query logs in different languages to improve multilingual content and search. 2010-2013. (Competitiveness Innovation Framework Program - Information and Communication Technologies Policy Support Program, n°250430)

SYNC3
SYNergetic Content Creation and Communication. A framework to structure, render more accessible and enable collaborative creation of the extensive user-provided content that is located in personal blogs and refers to running news issues. 2009-2012. (Information and Communication Technologies - Collaborative Project, n°231854)
Europeana v1.0
Successor network to the European Digital Library thematic network which created the EDL Foundation and Europeana prototype. 2009-2011. (eContentplus Program-Thematic Network, n° ECP-2008-DILI-558001)

EuropeanaConnect

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning. PASCAL is developing the expertise and scientific results that will help create new technologies such as intelligent interfaces and adaptive cognitive systems. 2008-2013. (Information and Communication Technologies - Network of Excellence, n°216886)

EERQI
European Educational Research Quality Indicators. 2008-2011. (SSH7-Collaborative project, n°217549)

PINVIEW
Personal Information Navigator Through Viewing. 2008-2010. (Information and Communication Technologies - Collaborative Project, n°216529)

SHAMAN

COST
Computational Social Choice. Network that addresses the fundamental new challenges in the design and analysis of methods for collective decision making raised by recent technological advances in social networks, electronic commerce, webpage ranking, and e-governance. 2012 - 2013. (ICT COST Action IC1205)

Framework Programme 6 (2002-2006)

CACAO

SAPIR

SMART

VIKEF

REVEAL THIS

THETIS
Training for Hotel Employees to Interact in Situation. Thetis focused on the creation and localization of enhanced on-line pedagogical content for language learning in the tourism industry. Jan-Dec 2004. (European digital content for the global networks-eContent, n°28655)

PASCAL

CONVIVIO

**LAVA**

**TransType II**

**MILK**

**KerMIT**

**MUCHMORE**

**PIE**
Progress Instance Evolution. The investigation & demonstration of what “services” are needed to support human intensive processes that are typically long lived, distributed and always submitted to dynamic evolution to cope with real industrial and market pressure. 1999-2001 (Long Term Research ESPRIT IV project n°34840)

**TRINDI**
Task-Oriented Instructional Dialogues between Man and Machine. Dialogues between humans and machines that enable the human to make choices in the performance of a certain task. 1998-2000 (Language Engineering Telematics Applications Programme n°IE4-8314)

CAMPIELLO
Enable people to interact and co-operate in building a new, richer sense of community based on exchange of knowledge sedimented over cultural resources. 1997-2000. (IST Esprit Project 25572)

COORDINA
Collaboration in assessing and improving the various models and languages, in order to advance the state-of-the-art. Also, put to the test the theoretical advances by studying test cases based on real applications. 1997-2000. (Applied Research and Development Telematic Systems n°24.512)

EUROgatherer
Designing and implementing a system which provides a personalized information gathering service and is based on software agents. 1998-2000. (Telematics Information Engineering n°IE-8011)

**TwentyOne**

STEEL

SPARKLE

GLOSSER
Helping people who partially know a language but cannot read it quickly. 1995-1997 Copernicus

COMPASS
Adapting bilingual dictionaries for on-line COMPrehension ASSistance. Implementing two bilingual dictionaries (English-French and German-English) in an on-line context sensitive comprehension dictionary. 1994-1996. (Linguistic Research and Engineering. n°LRE62080)


**TMR**
Training & Mobility of Researchers in Machine Learning. 1997-2002 (Research Network Learning Computational Grammars, n°ERBFMRX-CT98-0237)

**CAMPIELLO**
Enable people to interact and co-operate in building a new, richer sense of community based on exchange of knowledge sedimented over cultural resources. 1997-2000. (IST Esprit Project 25572)

**COORDINA**
Collaboration in assessing and improving the various models and languages, in order to advance the state-of-the-art. Also, put to the test the theoretical advances by studying test cases based on real applications. 1997-2000. (Applied Research and Development Telematic Systems n°24.512)

**EUROgatherer**
Designing and implementing a system which provides a personalized information gathering service and is based on software agents. 1998-2000. (Telematics Information Engineering n°IE-8011)

**TwentyOne**

**STEEL**

**SPARKLE**

**GLOSSER**
Helping people who partially know a language but cannot read it quickly. 1995-1997 Copernicus

**COMPASS**
Adapting bilingual dictionaries for on-line COMPrehension ASSistance. Implementing two bilingual dictionaries (English-French and German-English) in an on-line context sensitive comprehension dictionary. 1994-1996. (Linguistic Research and Engineering. n°LRE62080)
French Government Projects

**FIRE-ID**
Fine-grained analysis and recognition of photos or videos shared on social networks, scanned documents or images captured for video surveillance. 2012-2015 (ANR-CONTINT, n°2012 CORD 016 02)

**IMAGIWEB**
Analysis of the image life cycle through the Web 2.0. 2012-2015 (ANR-CONTINT, n°2012 CORD 002 06)

**DYNAMICITE**
Data analytics and visualization to create a city wide view. Using traditional and official data about a city but also analytics of spontaneous conversations on the web. 2012-2014 (CDC-FSN, n°013246-337863)

**CHRONOLINES**
Creating event-based chronologies in a semi-automatic way by analysing news articles in multiple languages. 2011-2014 (ANR-CONTINT, n°2010 CORD 010 03)

**COCLICO**
Creating new dynamics for the French software forge communities by giving a framework to this free ecosystem. Project labelled by Minalogic and Systematic Competitiveness Clusters. 2009-2011 (Ministère de l’Economie, de l’Industrie et de l’Emploi-DGCIS-STIC, n°09.2.93.0233)

**ALADIN**
Advanced text mining of patient records to help control Hospital Acquired Infections (HAIs). 2009-2011 (ANR-TecSan, n°08-TECS-001-01)

**TANGUY**
A system to extract “micro-knowledge” from texts and stock it in a semantic network. 2008-2012 (ANR-CONTINT, n°2008 CORD 022 02)

**OMNIA**
Analysing multimedia documents in 3 dimensions (image, text, emotion) and in a multilingual context. 2008-2010 (ANR-MDCO, n°07 MDCO 009 01)

**Atash**
Designing learning algorithms for tree transformations and their implementation for the integration of document data (PDF, html, doc) in XML databases. 2005-2009 (ANR-RNTL, n°05RNTL00102)

**INFOMAGIC**
Creation of robust, easy-to-use information extraction and new knowledge consolidation tools to increase analysis and decision-making capabilities of teachers, competitive watchers, financial analysts, military personnel and researchers. (CAP DIGITAL). 2007-2009 (Ministère de l’Economie, des Finances et de l’Emploi-SDGE-STIC, n°07 2 93 0763)

**SAFIR**

**TRINDI**
Task-Oriented Instructional Dialogues between Man and Machine. Dialogues between humans and machines that enable the human to make choices in the performance of a certain task. 1998-2000 (Language Engineering Telematics Applications Programme n°LE4-8314)
Imaging Demos

Analyse, classify & retrieve in new, creative ways
Launched in 2011 “Open Xerox” is an online technology portal that provides instant access to the latest Xerox innovations. Anyone can test pilot technology and provide feedback directly to lab scientists and engineers.

The technologies featured are early or incremental releases that have not gone through the rigorous testing and validation generally conducted by product testing groups.

There are two main categories of technologies on the portal. The first, SOAP or RESTful Web services, are accompanied with API documentation to help you access them. The second group are web applications that run on the portal server. For your protection, Open Xerox incorporates security and full access rights management. Single Sign On transfers user credentials to all apps and services.

Open Xerox is also a gateway to information on other interesting Xerox lab projects, or to (open source) software that can be downloaded and executed on a user desktop, with instructions user manuals and starter kits. These kits are accompanied by appropriate licensing agreements.

www.xrce.xerox.com

Open Xerox
Analyse, classify & retrieve in new, creative ways
Acknowledgments

We sincerely thank Suzette Norris for editing the anniversary articles, the Xerox Research Centre Europe director Monica Beltrametti, laboratory and area managers for their contributions and in particular Chris Dance for his help with numerous sections, Florence Helly for the Open and Collaborative Innovation section, Tom Zell for Patents and all the authors of the anniversary articles, Linda Emberley for her InDesign help and the Xerox branding team for their support.

Irene Maxwell and Claudia Heyer
Xerox Research Centre Europe