MLi Research Observatory: WP5

Hamish Cunningham
Professor of Computer Science
University of Sheffield

December 10, 2014

http://mli-project.eu/
https://gate.ac.uk/
https://hamish.gate.ac.uk/

Summary

- Mission: highlight the potential of infrastructure in promoting tech transfer roll-over for EU language and knowledge industries
- Method: digging into the startup and SME scene
- Outcome: recommendations for MLi and for future EC interventions (including the long-term and the blue sky!)

The Terrain

- Social Media
  - problems of access and challenges of scale:
    - Twitter: DataSift and NTT Data only independents with firehose access (Apple bought Topsy, Twitter bought GNIP)
    - Facebook: a closed book
    - The Rest: a long tail
  - the four Vs:
    - recent focus: volume, variety, velocity
    - upcoming: veracity
- Big Data
  - infrastructure:
    - leading player still AWS (Amazon Web Services)
    - some inroads by App Engine, Azure, Rackspace & smaller players
    - open alternatives (OpenStack, CloudStack, Eucalyptus, OpenNebula, ...) reaching towards critical mass
  - analytics startups: throw a rock
- Problematic issues:
  - privacy and security in the age of mass surveillance (below)
  - predominant model: advertising (zero contribution to inclusion, resilience, quality of life)

EU Industry: Contradictory Pressures

- majority of EU LT suppliers in the small-to-medium size range
- social/mobile creating contradictory forces:
  - constriction of the market spaces for small players due to:
    - data volume rises and infrastructure costs commensurately higher;
    - the economies of scale open to the big players become proportionately greater as the gap between large and small increases
  - new opportunities for innovative first movers, but only available to those able to exploit 3rd-party infrastructure for scaling without prohibitive cost
    - distribution of these abilities currently skewed towards the north-western and transatlantic technology communities

EU Industry: Contradictory Pressures (2)

Specific examples of operations that are currently prohibitively expensive for European SMEs and research labs:

- the filtering and supply of large web fragments (for example all the new or changed pages in France yesterday, or all the regularly changing commercial pages on `.com`, or all the pages with high page-rank, or etc.)
- the filtering and aggregation of real-time social data streams (for a promising — and European — filtering example see DataSift.com)
- time series analyses or geographic aggregation of opinion or sentiment (it it was impossible last year, for example, to get a broad measure of European thinking with respect to the Greek situation, as a leading Athens text processing researcher pointed out to me recently)

A European Language Data Success Story...

Technology transfer for *Compact Predictive Typing* on touch devices

- when the inventors of a leading Android predictive typing system won time on CERN's Large Hadron Collider they used it to crawl as much fluent multilingual text as they could
- the result: a successful business that employs 100 people in London
- they took the massive volume of language data that they were able to collect on CERN's machines and built statistical models
- encoded them in a very compact format (using approximate hashing)
- use them to predict the next word during typing on smart phones and other touch-based devices

MLi / **WP5**: how to best promote more stories like this? (via a. requirements on the MLi Hub architecture(s), and b. via programmatic interventions for e.g. CEF/AT)

Success Story (2)

Lessons:

- it would have been impossible to achieve without access to an extremely large computational infrastructure: the data volumes are truly enormous (CERN Data Centre processes about one petabyte of data every day; hosts 10,000 servers with 90,000 processor cores)
- the computation was *bursty* — a classic case for cloud computing
- Google could have done it, but didn't

These suggest that:

- small players can still enter the market spaces of the giants in cases where the infrastructural requirements are not constant
- these are typically client-side with only intermittent large-scale computation
- this is *not* true of 24/7/365 applications like public search infrastructure — here the EC would need to make a different category of intervention (see below)

Preliminary Recommendations (1)

- Prioritise availability of extremely large infrastructure for language technology (see deliverable 5.1 section 2; also section 7).
- Prime the technology transfer pipeline with research and startup use cases (see section 3).
- Develop XaaS marketplaces for the deployment and monetisation of CEF building blocks (see section 4).
- In the specific case of translation (see section 5):
  - prioritise translation of user-generated content
  - incorporate crowd-sourced models of translation
- Taxonomise the technology transfer roll-over point (see section 6) as a foundation for infrastructure design.

(XaaS: IaaS, PaaS, SaaS...)

Preliminary Recommendations (2)

- *Reuse not reinvention*: leverage the success of AWS, OpenStack, Hadoop, S4, and the like.
- *Applications-driven*: infrastructural activity that is insufficiently bound to applications is a resource sink with little chance of long-term utility.
- Active support for decentralised social networks as a vector for privacy preservation and European diversity.
- Provisioning of the European Language Cloud.
- Network as societal infrastructure:
  - search and social media are the lifeblood of the digital economy
  - they should therefore be regarded as social infrastructure
  - we don’t expect our roads to make money — neither should we expect the basic functions of the network

Futures (1): the Device Frontier & LT Solutions

- LT was at the heart of the mobile revolution; continues as main conduit between users and on-line needs via
  - search, geolocation, translation and etc.
- beyond mobile, advances in low power chips (chiefly the EU's ARM) are spawning new generation of LT-powered devices
- e.g. Amazon Echo was announced last week — http://www.amazon.com/oc/echo/
  - combination of a small device doing speech recognition on the client and question answering from a KB of information extraction data in the cloud
- Europe has both LT expertise and a resurgent emedded hardware ecosystem (e.g. ARM, but also Arduino or Raspberry Pi)
- combination of these two strengths could drive innovative products and services meeting societal challenges
  - e.g. devices that help the elderly control smart homes, or address exclusion by increasing local community connectedness

Futures (2): Euroogle?! (a CERN for the Web)

If the EU parliament things we should break up Google, perhaps the time is ripe for a new research programme...

> *The web was invented at CERN, the European particle physics lab, and grew beyond all precedent to be the centerpiece of the information revolution... and just as we need to understand the physics of our world to prepare for our future needs so we also must understand and profit from the web.*
>
> *We live in times when the European ethos of a strong and stable civil society with high cohesion and excellent social services must weather a tempestuous financial storm. It is more vital than ever to promote scientific and technological leadership in order to drive industrial growth and social progress. We need a **CERN for the web**!*

What does that mean?

A CERN for the Language Web (2)

- provision of a shared infrastructure for web R&D that can scale to petabytes and is open and flexible enough to cater for scientists and startups, governments and libraries, technologists and citizens (perhaps building on the bare metal layers from e.g. http://www.helix-nebula.eu/; note new GATE work with Science and Technologies Facilities Council)
- transition of research programmes in language technology onto the shared infrastructure, and the creation of a market for web analysis solutions, & foundation of an R&D programme on open search for Europe
- create new marketplaces for services on top of the cloud infrastructure (e.g. customer relations no longer sitting next to the phone; now reading 250,000 tweets per week)

The cost of an alternative to Google may be too large to swallow in one go; a smaller set of chunks can draw the path to a **solution** in the medium term.

Futures (3): Privacy in the Age of Surveillance

- Edward Snowden and others have exposed how the NSA (and their Five Eyes collaborators in the UK/Canada/Australia/NZ) is attempting to record and analyse *all* electronic communication, *everywhere*, *all of the time*
- to do this they actively subvert security systems by
  - tapping cables (between countries or between data centers)
  - weakenning standards (e.g. NIST eliptic curve crypto)
  - coercing telecoms and internet companies *in secret* (gagging orders)
- regardless of if this is a good idea it inevitably compromises both online privacy and online security
- closed systems must be assumed compromised by default
  - heartbleed and shellshock: bugs in open systems that have been fixed
  - closed systems shrouded in secrecy and open to coerced backdoor insertion

Links

- these slides:
  - HTML: https://hamish.gate.ac.uk/pages/about/talks/mli-review1/
  - PDF: https://hamish.gate.ac.uk/pages/about/talks/mli-review1/index.pdf
- MLi: http://mli-project.eu/
  - WP5 deliverables: http://mli-project.eu/?p=490
- GATE: https://gate.ac.uk/
- me: https://hamish.gate.ac.uk/