The Gistr Platform

Sébastien Lerique*

Abstract

Here goes the abstract.

Introduction

1 Web and Smartphone experiments

As an aside to online data collection on social networks and blogs, making experiments using the web and smartphones is really promising

1.1 The need for embedding

Some studies are quantitative, but lack a fine-grained view of what's a work.

Embedding means getting access to the way things are meant in the lives of subjects, it means getting access to context. Pick up all the metadata you want.

It lets you try avoid the quantitative vs. qualitative problem described by Becker.

1.2 Pros and cons

1.2.1 Pros

Just like controlled lab experiments, the general setting and framing of such a study is controlled.

As explained above, it's real-life embedded.

If needed, it can be massively interactive: any number of people interacting at the same time.

Recruitment is super flexible: big pool with filters, pay them or make it gamified.

1.2.2 Cons

Big engineering efforts, new techniques and technologies (async, user management, . . .).

No face-to-face control, so need for real anti-spam pressures.

*Centre d'Analyse et de Mathématique Sociales (UMR 8557, EHESS, Paris), and Centre Marc Bloch (Berlin). Email: sebastien.lerique@normalesup.org.

1.3 An example: the Daydreaming app

Successful app that took us a while to develop, but lets you answer questions that cannot exist in the lab.

2 The Gistr Platform

2.1 Presentation

As part of Sébastien's PhD thesis we are studying the transformation of short sentences – such as quotations from politicians or spokespeople – as they are propagated through various media. Our first study focused on the evolution of such short quotations as they are copied from blog to media outlet to blog. Indeed, authors often transform quotations when publishing them online despite the implicit and common-sense injunction to quote people verbatim: a few words disappear, a contraction appears, the quote is cropped, etc. (Simmons, Adamic, and Adar 2011; Lauf, Valette, and Khouas 2013). Given this observation, the data collected by Leskovec, Backstrom, and Kleinberg (2009) is at first sight a very good candidate to study the evolution of online content as it is transformed by users. But the actual study proved challenging for two fundamental reasons:

- *Missing information*: most blog and media outlet authors do not quote their source when they publish a quote online (it's often not relevant to the article), meaning there are no source-destination links in the data collected; this information must instead be inferred anew to be able to study the evolution of content. Equally limiting, there is no access to author information (gender or age, experience in writing, but also psychological factors like memory span).
- *Missing context*: the lack of access to the context of production and reception of quotes makes it impossible to interpret what a quotation means to its author or its reader (Wittgenstein 2010; Briggs 1992; Cuffari, Di Paolo, and De Jaegher 2014). Analysing any kind of semantic evolution is therefore out of reach for passively collected online data.

The Gistr platform emerged from a concern to address these two problems. The general goal is therefore the study of interpretation and sense-making of short sentences in particular contexts, and the effects the accumulation of these have at the global scale. How (linguistic) information is interpreted, that is, made sense of by people.

In short: **"What sense is made of what when, and what follows from that at large scale."**

Interpretation is a kind of attention/perception, that is exploration of an environment. It's a process of exploring information (through educated attention) and putting it to use in the current task. This involves many different processing levels, and is studied by several different disciplines. Here are a few examples of what is involved:

- Interactions: relating to people involves several tasks; e.g. trying to look good to someone (based on what you believe that person thinks) is a task that will influence how you perceive what they say. For social anthropologists, meaning is a shared property that emerges from the interaction.
- Cognitive biases in general: for some tasks we often make recurrent and predictable mistakes or transformations (e.g. word frequency effect [Yonelinas, 2002], cognitive dissonance [Festinger, 1962], risk perception [Kahneman & Tversky, 1996] [TODO: read], and any number of cognitive-, memory-, or heuristics-induced biases), because of low-level processes involved in the task.
- Metaphors: reasoning and idea-association are closely related to basic metaphors that pervade everyday language (like 'more' is 'up', 'later' is 'forward', etc.).

2.1.1 State of the Art

Up to now Epidemiology of Representations has focused on cultural bodily practices with long intergenerational lifecycles like

- smoking [Claidière & Sperber, 2007],
- bloodletting [Milton, Claidière, & Mercier, 2015],
- portraits [Morin, 2013], [TODO: read]
- religion [Boyer, 2001], [TODO: read]

interfacing with studies of diachronic evolution of language on the way [Claidière, Smith, Kirby, & Fagot, 2014]. [TODO: read]

Recently, practices with short intragenerational lifecycles like

- music [MacCallum, Mauch, Burt, & Leroi, 2012], [TODO: read]
- risk perception (abstracting away from individual variation) [Moussaïd, Brighton, & Gaissmaier, 2015 - in press], [TODO: read]

that have less to do with changes in bodily practice, and more with interpretation, have started to be studied. I aim to bring a new case in this area of short lifecycle opinion dynamics by studying the semantic evolution of short sentences and short stories in interpretation chains. What change takes place here is mainly due to interpretation and the reconstructional component of memory which, as mentioned above, involves many levels and is influenced by many factors. Therefore, after starting at the macro scale where individual variation and context details are abstracted out, I also aim to gradually move towards the meso scale, integrating more contextual and personal details and factors as I go along.

Finally, even if it's not my main target, this study should also be of interest to the big data/Twitter/Facebook studies community, where well-controlled data is hard to come by (if you don't work for Facebook).

2.2 Criticisms

Other disciplines work on this subject with different focuses and valuable criticisms. The three areas I roughly identify are:

2.2.1 Linguistics

I'm not a linguist, and this study is obviously affected by linguistics-level effects that I don't know about. I plan to address that issue, in a limited way, by augmenting my basic knowledge of linguistic effects by going to the LOT Summer School 2015, presenting my stuff at Martin Hilpert's discussion group "Issues in language variation and change", and asking around as much as I can.

2.2.2 Situated/embodied/extended cognition

Epidemiology of Representations rests on the fodorian idea of cognition where the brain is a storehouse of representations and everything happens in there. During the 2000's there has been a substantial move in philosophy of mind from this position to a position positing the central role of environment. In a nutshell, cognition is an exploration of the environment (vs. storing representations of the environment) and as such cannot be considered out of its environment. Some philosophers go so far as to consider that the environment is part of the cognitive device.

This line of thinking has greatly improved the way cognitive science analyzes contextual/situational information and interaction, has led to many experimental breakthroughs (that I know of, mostly in perception, e.g. perception substitution devices [O'Regan & Noë, 2001] [TODO: read]), and has a number of implications for how perception and affect should be tackled (e.g. Bower & Gallagher, 2013 [TODO: read]).

2.2.3 Social anthropology

Adopting much of the "embodied" position w.r.t. representations, social anthropologist Tim Ingold criticizes Sperber's approach to cultural evolution on a number of points which can be summed up as follows:

- Culture is not something that is *transmitted*, but something that one *grows into* (Ingold, 1997). Hence the driving force of cultural change is not how representations are transmitted, but how attention is educated. As such, cultural and natural evolution are one and the same process, and separating them into one taking place before the other (Ingold, 2004) only hinders the analysis ("Cultural differences, in short, are not added on to a substrate of biological universals; rather they are themselves biological").
- Meaning (and therefore interpretation) is not a property of representations, it is what emerges from the individual's history with a particular representation; interpretation is how a piece of information is put to use in one's life (Briggs, 2006).
- An alternative to studying culture as an evolutionary process parallel to genetic evolution is to consider the whole ecosystem as a field of developmental systems (Ingold, 1998).

2.2.4 Treatment

Each of these disciplines would have many valuable criticisms addressing the shortcomings of how this study is currently conceptualized and planned. But as explained above, the plan is to start from the Epidemiology of Representations approach and, if time allows, gradually incorporate criticism as it serves the purpose of explaining the collected data and refining the conditions. (For instance, the embodied and social anthropology critique would prove useful in trying to develop conditions taking the context into account, which in turn would probably allow us to explain some noise.)

The bottom line is that all studies have their shortcomings and these criticisms will most likely not be addressed in the immediate future. Although the distant future sure would like to.

2.3 Breakout and development

2.3.1 Breakout

General goal: explore some interpretation effects, at the single and cumulative levels, in tasks involving sentence and story rewriting or reformulation. Interpretation involves many levels of complexity, so we start with the simplest possible condition (probably underspecified), and add new measures and conditions as we go to see if we can explain our noise.

To do this, we build an experiment where subjects repeatedly memorize and rewrite (i.e. interpret and reconstruct) textual content in various situations and tasks. This will be:

- Content we selected for the experiment
- Content newly suggested by the subjects

- Content already interpreted and transformed by previous subjects; this lets us generate interpretation chains, with each piece of initial content generating a tree of successive interpretations.

Measures [TODO: add references to state-of-the-art]

- How many (and which) elements are maintained (manually coded)
- LDA to see the evolution of subjects (either on each tree, or on the whole corpus)
- Author categorization based on their LDA bias and element creation/loss. Correlations to demographic measures (age, gender, CSP).

Perspectives

- Effect of stimulus degradation, e.g. change of direction in the bias [FIXME: ref]
- Effect of social proximity [Scherer & Cho, 2003; Binder, Scheufele, Brossard, & Gunther, 2011] [TODO: read]
- Effect of emotional state [Kramer, Guillory, & Hancock, 2014] [TODO: read]
- Effect of interaction setup/framing [Kahneman & Tversky, 1996, for framing] [FIXME: embodied ref for interaction?]

2.3.2 Development

The experiment presents itself as an online Game With a Purpose where people can freely register and participate. The gamification aspects are likely to come later, once the experimental aspects are stable and implemented. This lets us start experimenting on paid (Prolific Academic) or volunteer (Crowd Crafting) platforms, and later open up to platform-free participation.

Technically, it's a web application that runs in the browser and can also be packaged as a native app for Android and FirefoxOS. The backend to this app is spreadr: it receives, stores, and dispatches content to the app.

See the Milestones on GitHub for the detailed planned development.

References

Briggs, Jean L. 1992. 'Mazes of Meaning: How a Child and a Culture Create Each Other'. *New Directions for Child and Adolescent Development* 1992 (58): 25–49. doi:10.1002/cd.23219925804.

Cuffari, Elena Clare, Ezequiel Di Paolo, and Hanne De Jaegher. 2014. 'From Participatory Sense-Making to Language: There and Back Again - Springer'. doi:10.1007/s11097-014-9404-9.

Lauf, Aurelien, Mathieu Valette, and Leila Khouas. 2013. 'Analyzing Variation Patterns In Quotes Over Time'. *Research in Computing Science* 70: 223–32. http://www.micai.org/rcs/2013_70/Analyzing%20Variation%20Patterns%20In%20Quotes%20Over%20Time.html.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. 'Meme-Tracking and the Dynamics of the News Cycle'. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. KDD '09. New York, NY, USA: ACM. doi:10.1145/1557019.1557077.

Simmons, Matthew P., Lada A. Adamic, and Eytan Adar. 2011. 'Memes Online: Extracted, Subtracted, Injected, and Recollected'. In *Fifth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2836.

Wittgenstein, Ludwig. 2010. *Philosophical Investigations*. John Wiley & Sons.