Some thoughts on the relationship between computational linguistics and literary scholarship

Hannah Alpert-Abrams with Dan Garrette

At the Association of Computational Linguistics conference in Bulgaria last month, researchers from CMU presented a model for cinematic archetypes: "Learning Latent Personas of Film Characters." The model uses the descriptive language of Wikipedia entries along with personal data of actors in films to automatically induce a set of character personas: *the traitor, the flirt*.

As a literary scholar who studies both novels and films, I read the paper with interest, curious to see how NLP researchers could advance my field. The paper opens with a gloss of Aristotle: the debate over the supremacy of plot versus characters. It goes on to situate itself within what literary scholars call archetypal theory, referencing the anthropologist Joseph Campbell and the psychotherapist Carl Jung. I was pleased to see this nod towards several millennia of thought on the nature of literature.

I was disappointed, however, to notice that the literature review failed to mention any of the work done in the latter half of the twentieth century. This is a problem for the utility of the model from the perspective of literary scholarship. The paper is clearly exemplary NLP work, as evidenced by the fact that it was published at the most prestigious conference in the field. But to be truly interdisciplinary, the project must be cutting edge in the field of literary scholarship too.

A brief history of archetypal theory: I don't work in archetypes. That's because no one works in archetypes anymore. Jung and Campbell developed theories that attempted to find narratives and personas that were shared by all cultures. Although their work remains historically significant, the general consensus today is that Jung's *collective unconscious* and Campbell's *monomyth* are overly generalized concepts. The pursuit of a universal theory that could encompass the entirety of human experience tended to force all cultures into a Western European worldview. This has since been replaced by the recognition that human experience is culturally specific.

Northrop Frye, the scholar who most famously applied archetypal theories to literature, is sadly absent from this paper. But his work can help us understand why even a smaller, more culturally specific form of archetypal theory — what Bamman et al. refer to colloquially as stereotypes — has similarly fallen out of favor in literary study. Frye's project was modeling the meta-structures of Western literature to define basic literary tropes. Bamman et al.'s model is able to automatically induce the kinds of categories that interested Frye. The direct contribution of their model to archetypal theory is that it allows us to test our preconceived notions of stereotypical personas. Did we get it right? Did Northrop Frye?

But Frye's work, like that of Jung and Campbell, is only rarely cited in literary study today. This is because he was interested in building textual models, but did not consider the social and historical context in which these texts were produced. More interesting to contemporary scholars than the archetypes or stereotypes themselves is the way that they reflect systems of power and oppression.

When we look at Wikipedia entries about film, for example, we would not expect to find universal, latent character personas. This is because Wikipedia entries are not simply transcripts of films: they are written by a community that talks about film in a specific way. The authors are typically male, young, white, and educated[1]; their descriptive language is informed by their cultural context. In fact, to generalize from the language of this community is to make the same mistake as Campbell and Jung by treating the worldview of an empowered elite as representative of the world at large.

To build a model based on Wikipedia entries, then, is to build a model that reveals not how films work, but how a specific subcategory of the population talks about film.

By focusing on cinematic archetypes, Bamman et al.'s research misses the really exciting potential of their data. Studying Wikipedia entries gives us access into the ways that people talk about film, exploring both general patterns of discourse and points of unexpected divergence. We might wonder, for example, whether there are cases where a statistical model misidentifies a character, and why that misidentification occurred. If, for example, the model frequently mislabels female *villains* as *flirts*, it might suggest something about the way that Wikipedia writers think about women. More generally, we might wonder whether the race or gender of an actor skews the way that the character is represented on Wikipedia: a female hero might be described differently from a male hero, or a black hero from a white hero.

These are interesting questions, and as a literary scholar I find it exciting that there are technologies we can use to approach them. Sitting down to read 42,000 Wikipedia entries is prohibitively inefficient, and computational linguists are absolutely correct to think that their tools will be useful for the study of literature.

But literary scholarship can also inform computational linguistics, if it is fully incorporated into project development and analysis. The lack of regard for the work being done in the humanities was seen this summer in several high-profile articles on the relationship between science and the humanities, including Steven Pinker's *Science Is Not Your Enemy* and Lee Siegel's *Who Ruined the Humanities?* Siegel reduces literary study to the reading of great novels, while Pinker proposes a one-sided model through which science is the salvation of the humanities.

If computational linguistics is going to 'save' the humanities, it will do so by engaging with contemporary work already being done in the field. Had Bamman et al.'s paper been more thoroughly informed by literary theory, the authors would have started with more appropriate questions about their data and provided results that fit better into academic discourse about film. This kind of research could open new realms of possibility in both literary studies and computational linguistics.

The great thing is, to address these problems computational linguists don't have to tackle the *Norton Anthologies* on their own. There are already people who have built their careers on this kind of research. Just take one of us out to coffee.

[1] http://www.wikipediasurvey.org/docs/Wikipedia_Overview_15March2010-FINAL.pdf