Dan Cohen's Digital Humanities Blog » Blog Archive » Million Books Workshop Wrap-Up

May has been a month of travel for me (thus the light posting in this space). I gave a talk about Zotero[1] and related developments in the humanities and technology at the Stanford Humanities Center[2], and spoke at the annual meeting of the American Council of Learned Societies[3] about how digital research is a major emerging theme in scholarship. Finally, I participated in the Tufts[4] "Million Books" Workshop[5], which explored the technical feasibility and theoretical validity of extracting evidence and meaning from the large new corpora of online texts. The three main topics were how to get from scanned documents (especially the complicated ones that scholars sometimes encounter, like Sanskrit manuscripts or early modern broadsides, rather than simply formatted texts like modern English books) to machine-readable text that can be searched and analyzed; machine translation of texts; and moving from text to actionable data (e.g., extraction all of the place names from a document or summarizing large masses of text). Some developments worth noting from the workshop:

I had vaguely heard about the open-source optical character recognition (OCR) project OCRopus[6], but Thomas Breuel's[7] detailed description of the project made it seem extremely promising, especially for scholarly applications. Even after two decades of research and development, the error rate of OCR is still too high for many historical texts, and atrocious for compound texts like Victorian mathematical monographs[8] (with all of those equations that end up, improperly and disastrously, as regular text after OCR) or works with vertical text (e.g., Japanese poetry) or images. OCRopus ambitiously plans to support any language written in any direction with any layout. It also breaks down the conversion of scans to text into separate processes that produce *probabilities rather than certainties*. This is critical. Most OCR packages give you a text result without noting where the software was unsure of a word or letter.

Thus you might get "Cohem" rather than "Cohen" without knowing that the software thought long and hard about the correct interpretation of that last letter. OCRopus instead produces a statistical output that says to any end-user application (like search), "I'm sure about 'Cohe' but the last letter has a 60% probability of being an 'm' and a 40% probability of being an 'n'." A search for "Cohen" could thus return the document as a result even if the "final" transcription defaults to "Cohem."

OCRopus also uses far more sophisticated methods than current OCR software to find titles, ordered blocks of text (like columns), and marginalia. Brilliantly, rather than outputting XML at the end of its processes, OCRopus outputs to XHTML and CSS3 so that it can much more accurately represent the fonts and layout of the original. Very impressive. The project is just in pre-alpha right now with a 1.0 release to come in the fall of 2008. Unsurprisingly, OCRopus is supported by Google[9], which plans to use it for Google Book Search[10]. (Right now they have OCR that's good enough for search, which doesn't need anywhere near 100% accuracy, but they plan to re-OCR their book scans with OCRopus when it's ready.)

David Smith[11] spoke about the cutting edge of machine translation (i.e., the use of computational methods to translate text from one language to another). The field seems extremely active right now, and new methods promise better translations in the near future. David spoke of several developments. First, many projects are seeding their software with parallel texts, such as documents from the United Nations or the European Parliament, which are translated very precisely by humans into many languages. Parallel text corpora (with English as one of the parallels) on the order of 20-200 million words (roughly 1-10 million sentences) are available for a number of languages. Unfortunately, the parallel texts often come from genres like laws, parliamentary proceedings, and religious texts (not only the Bible but also, quite interestingly, *Dianetics* is one English text that has been translated into virtually every language, including Uzbek). These genres are, of course, less than optimal for widespread translation uses. We might, however, be

able to use parallel translated works from Google's scans or the Open Content Alliance[12] to help improve the seed corpus.

Second, David noted the resilience of n-gram[13] analysis—breaking down a document into word pairs or triads. Usually you can predict the next word in a document by looking at the previous two words and then assessing the probability of the word following each pair. Most of the best machine translation services (like Google's) now split a text into bi-grams and tri-grams (two- and three-word pieces) and then translate those n-grams into very exact parallels in the target text using an n-gram library. This is better at keeping the style of the text and avoiding the off-sounding literal translations that have dogged the field. David feels that machine translation has reached the point where it can very usefully tell a user when a primary source document has been mistranslated by a human, which can be very useful for scholarship.

Finally, David Mimno[14] discussed how to move from the text that results from the work of OCR and machine translation (if necessary) into forms that will help with research and analysis in the humanities. David has been doing impressive work in document classification, i.e., computationally assessing a set of digitized texts and figuring out which ones are letters or poems or lab notes, or if the documents are all articles, separating them out into topic clusters. Like machine translation and OCR, when you begin to look under the hood this is an extraordinarily complicated field. The three main techniques—support vector machines[15] (SVM), naive Bayes[16] (probably the best-known method, often used in spam filters), and logistic regression[17]—are best viewed mathematically, and so lie beyond the scope of this blog. David is working on the Mallet project[18] at the University of Massachusetts, Amherst, which seems promising for document classification (a topic we are increasingly interested in at the Center for History and New Media[19] for historical research). The software is still in alpha but I plan to keep an eye on it.

Obviously a lot to think about from the month of May. How do we get

these complicated tools to scholars who don't have technical skills? How can we use these tools to reveal new, meaningful information about the past, without reproducing the obvious using computational means[20]? As I felt at the National Endowment for the Humanities meeting in April[21], the application of digital methods to the humanities is experiencing a burst of energy and attention in 2007. It will be interesting to see what happens next.

This entry was posted on Thursday, May 24th, 2007 at 10:10 am and is filed under Text Mining[22]. You can follow any responses to this entry through the RSS 2.0[23] feed. You can leave a response[24], or trackback[25] from your own site.

References

1. ^ Zotero (www.zotero.org)
2. ^ Stanford Humanities Center (shc.stanford.edu)
3. ^ American Council of Learned Societies (www.acls.org)
4. ^ Tufts (www.tufts.edu)
5. ^ "Million Books" Workshop (devwiki.perseus.tufts.edu)
6. ^ OCRopus (www.ocropus.org)
7. ^ Thomas Breuel's (www.iupr.org)
8. ^ Victorian mathematical monographs (www.dancohen.org)
9. ^ Google (www.google.com)
10. ^ Google Book Search (books.google.com)
11. ^ David Smith (www.cs.jhu.edu)
12. ^ Open Content Alliance (www.opencontentalliance.org)
13. ^ n-gram (en.wikipedia.org)
14. ^ David Mimno (www.cs.umass.edu)
15. ^ support vector machines (en.wikipedia.org)
16. ^ naive Bayes (en.wikipedia.org)
17. ^ logistic regression (en.wikipedia.org)
18. ^ Mallet project (mallet.cs.umass.edu)
19. ^ Center for History and New Media (chnm.gmu.edu)

20. ^ without reproducing the obvious using computational means (www.dancohen.org)
21. ^ National Endowment for the Humanities meeting in April (www.dancohen.org)
22. ^ View all posts in Text Mining (www.dancohen.org)
23. ^ RSS 2.0 (www.dancohen.org)
24. ^ leave a response (www.dancohen.org)
25. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen's Digital Humanities Blog » Blog Archive » Million Books Workshop Wrap-up*

http://www.dancohen.org/2007/05/24/million-books-workshop-wrap-up/

Readability — An Arc90 Laboratory Experiment

http://lab.arc90.com/experiments/readability