Towards an Automatic Index Generation Tool

Patrick Juola *(juola@mathcs.duq.edu)*
Duquesne University

Almost every non-fiction author has been faced from time to time with the generation of an index. Most novice authors (myself included) are taken aback by the magnitude of the task and the limited amount of computational and software support available.

The current state of the art is significantly improved from the days of 3-by-5 'index cards', (a telling term?), but only in mechanical, not intellectual terms. Modern publishing practice typically involves the author delivering a machine-readable 'manuscript', written in a document-processing system such as *LaTeX*. Index entries are defined as specific term/location pairs by the author. For example, an index entry written in *LaTeX*, might look as follows

```
The \index{Pittsburgh!University of}
University of Pittsburgh was established
in \index{Pittsburgh!city of} Pittsburgh,
Pennsylvania, in the year....
```

This will create an index entry on the 'current page', under the heading "Pittsburgh, University of" (as opposed to "Pittsburgh, city of," which would be the second entry, a related but separate subentry). Although guidelines for a good index (Northrup; University of Chicago Press Staff) are commonly available, the process of producing a good index is still largely unsupported, even by major and relatively sophisticated publishing companies such as Prentice-Hall.

What differentiates an index from a mere concordance?

There are at least six cognitive tasks (Maislin; Saranchuk) related to the production of a good index, as follows. Current standard support covers only the last.

- Identification of terms to index;
- Location of all informative references in the text;
- Identification/location of synonymous terms (e.g. "University of Pittsburgh" / "Pitt");
- Splitting of index terms to split into subterms;
- Development of cross-references within the index itself;
- Compilation of page numbers,

I will present a framework for the development of a 'machine-aided index generation system'. This bears the same relationship to an automatic indexer that machine-aided translation (MAT) does to machine translation (MT), in that it provides suggestions and reduces the overall workload for the human, but post-editing will still be necessary. Specifically, recent results in corpus linguistics (Charniak; Manning & Schuetze), including the development of taggers for part of speech (Cutting et al.; Schmid) the availability of ontologies and semantics networks, plus the light semantic analysis capabilities of latent semantic analysis (Landauer et al.), can be combined in a multi-phased iterative framework and implemented as user-level software. This paper presents some aspects of "good" indices (Northrup; University of Chicago Press Staff) and illustrates how they can be achieved computationally.

In general, following the University of Chicago's dictum that "it is always easier to drop entries than to add them, err on the side of inclusiveness," (rule 18.120) we start by assuming that every term is a potential index entry and look for criteria by which to eliminate enough terms to produce a reasonably-sized index. (5-15 references/page, between 2% and 5% of the length of the final work, according to rule 18.120.) For example, rule 18.8 states that "the main heading of an index entry is normally a noun or noun phrase---the name of a person, a place, and object, or an abstraction." A first pass, then, can use the results of a part-of-speech tagger and eliminate all terms that do not appear as a noun in the document. Within this set of nouns, I suggest two possible heuristics for further pruning; first, common nouns that are too common or too rare are unlikely to be useful index terms, and second, words that are too uniformly distributed are unlikely to be useful index terms. On the other hand, a case can be made that all proper nouns should be included. Other suggested heuristic will be discussed.

Within a single index term, "an entry that requires more than five or six locators is usually broken up into subentries" (rule 18.9). This can be treated as an example of word-sense disambiguation, for example, between *Pittsburgh* (University of) and *Pittsburgh* (city of). Again, I conjecture (and present supporting evidence) that existing technology can provide a useful and helpful basis for later human editing. Specifically, existing semantic representation techniques can model the context, and therefore the meaning, of each index token. For truly polysemous terms, cluster analysis of the set of token representations should yield a set of clusters equivalent to the degree of polysemy; by setting the separation threshhold to an appropriate level, the analysis can be forced to produce clusters of maximum size at most 5-6. At the same time, passing and uninformative references can be expected to produce isolated 'clusters' containing a single outlier — a strong candidate for omission. Once a list of index terms is collected, tokens not on that list can be compared in their semantic representation for

similarity with existing index terms; any word with near-identical meaning is a potential synonym and a candidate for a cross-reference.

Unfortunately, the evidence to be presented is largely heuristic and exploratory in nature. We are currently developing a prototype system, using LSA (Landauer et al.) and elementary corpus statistics such as TF-IDF to identify index terms. We also have a well-developed and intuitive GUI wizard for ease of use by a non-technical user. At present, the planned heuristics may or may not be sufficiently reliable to use without a human post-editor. However, if they can be shown to substantially reduce the work load on the human author, the resulting tool may still be of interest. I present the results of some prototype-scale experiments, plus some ideas about usability testing and the directions of future development.

Bibliography

Charniak, E. *Statistical Language Learning.* Cambridge, MA: MIT Press, 1993.

Cutting, D., J. Kupiec, J. Pedersen., and P. Sibun. "A practical part-of-speech tagger." *Proceedings of the Third Conference on Applied Natural Language Processing.* Trento, Italy, 1992. 42-46. Association for Computational Linguistics. Also available as Xerox PARC technical report SSL-92-01.

Landauer, T., P. Foltz, and D. Laham. "Introduction to latent semantic analysis." *Discourse Processes* 25 (1998): 259-284.

Maislin, S. "The cognitive half of indexing." *Proceedings of Massachusetts Society of Indexers Fall Conference.* Massachusetts Society of Indexers, 1996. n. pag. Association for Computational Linguistics. Also available as Xerox PARC technical report SSL-92-01.

Manning, C., and H. Schuetze. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press, 1999.

Northrup, M.J. "The role of indexing in technical communication." *Proceedings of SIGDOC-90.* Association for Computing Machinery, 1990. n. pag.

Saranchuk, G.R.Z. *A new index for the graduate program manual of the Faculty of Graduate Studies and Research at the University of Alberta.* University of Alberta, 1996. Accessed 2005-04-11. <http://www.slis.ualberta.ca/cap03/georgina/lis600fgsr_introduction.htm>

Schmid, H. "Part-of-speech tagging with neural networks." *Proceedings of SIGDOC-90.* Proceedings of COLING-94, 1994. n. pag.

University of Chicago Press Staff. *The Chicago Manual of Style.* 15th ed. Chicago: University of Chicago Press, 2003.