Information Theater versus Information Refinery

Doug Cutting **Per-Kristian Halvorsen** **Jan Pedersen**
Meg Withgott

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, California

Where Intelligence Resides

Intelligent behavior of a system is often equated with its ability to perform logical or heuristic inferences. Intelligence is thought to reside in the ability to *refine* the knowledge base, and thus in the articulation of "new" facts beyond the set of assertions explicitly represented. The fragility and complexity of the inference mechanisms required for this invite attempts at imbuing other components of the system with "intelligence".

We have chosen to focus on the intelligence required to locate knowledge instead of that required to infer facts. The problem we therefore start with is: what should one *present* in response to a query, and what sort of information can be *reliably* extracted from running text? Consider a text-based intelligent system that does not refine text into something resembling facts, but rather presents and possibly annotates interesting text passages for perusal by the user. This distinction between fact-inference and text presentation means replacing the requirement for deep understanding by the more feasible goal of text selection. It also implies a certain degree of forgiveness: the system is not required to be perfectly accurate since a person ultimately assists in the interpretation.

Pursuing systems that are capable of dealing intelligently with texts demands consideration of (1) the role of the user; (2) scale of data requirements; and (3) the interplay between general and specially-tailored processing techniques. We will discuss each of these considerations, and set forth a vision of the "information theater," drawing on our experience in implementing systems for intelligent information access.

Information Theater

The end product of information theater is texts or fragments of text arranged in a way meaningful to a user. High-bandwidth interaction greatly influences the nature of the underlying search and analysis algorithms. This has been brought out in our experience with the development of a text-access paradigm that weaves together interface, presentation and search in a mutually reinforcing fashion [3, 1]. Second, because text originates in paper as well as in electronic form, the documents may have physical attributes which are as important as their content. Consequently, as much as possible of the original appearance of the document is to be retained, including layout structure and relative positioning of text-pieces.

Third, in order for information theater to provide significant benefits beyond what is offered by, say a collection of books, its reach must include a vast and varied range of material. Indeed, small scale corpora are uninteresting from our perspective, as are methodologies that do not scale gracefully. In keeping with this view, some of our colleagues have constructed an interoperable substrate layer for storage and retrieval capable of binding together large and diverse corpora [12].

These considerations have lead us to adopt an architecture which interleaves statistical analyses, including new methods for document clustering [13], with coarse-grained content analysis [2] and selectively deployed, fine-grained parsing and semantic interpretation techniques [11],[4],[7]. This architecture permits the organization of the text collection along various generic, but interesting, dimensions, such as:

- Topic identification: Which documents cluster together?
- Stance: Is the document for or against a particular position?
- Focus: What parts of the document are most emphasized as determined by both textual and textural (or "linguistic and visual") cues?

By retaining information not only about the content of the texts, but also about their form and context of use, additional information is associated with the document:

- Author: Who wrote it and when?
- Origin: Factors with impact on the reliability of the information (Does the document come from USENET or a state law book?)
- Bibliometrics: What other text passages are referred to?

Thus a mixture of standard database techniques with statistical and linguistic methods provides the basis for presenting relevant text to the user in revealing ways.

Methodology

Statistical techniques are effective at determining the general topic of passages, and linguistic analysis may be able to identify other properties of a text passage, such as its perspective. Given such complementary abilities, we suggest that a user's request proceed through a series of computations intermixing fast filtration procedures with highly directed linguistic analysis.

Statistical Methods

If looked upon purely as data, text appears as a sequence of opaque tokens organized into sentences, paragraphs, sections, documents, etc. Patterns implicit in this data may be extracted through a battery of statistical methods that operate at varying granularities. For example, a standard lexicon and a large training set of text may be fed to a hidden Markov modeling procedure to yield a part-of-speech prediction rule that runs in linear time [9]. Traditional IR techniques, such as relevance feedback, relate documents by matching word-token distributions.

These sorts of adaptive, inferential methods are far from exhausted. Recent work at IBM Watson Research suggests that Markov modeling techniques may be applicable to the problem of reducing ambiguity in syntactic parsing by capturing context as probability statements [5]. Recent experiments we have carried out suggest that innovative agglomerative clustering algorithms can be quite effective at identifying groups of texts that fall within the same topic [13].

Statistical methods are only effective when informed of structural features unique to natural language discourse. In other words, drawing a strong distinction between statistical and NLP techniques is counter productive; each benefits from the other. For example, conventional statistical measures of similarity operate under the assumption that word tokens have unambiguous meaning and that text units (i.e. documents) have topical integrity. These assumptions are, of course, not true in detail. However, correlating a mass of data enables one to pick out a "signal" from the background "noise". Substantial improvements may be obtainable by reducing the "noise". If word tokens can be pooled into equivalence classes through an appropriate use of morphology, inflected and some derived forms can be counted as instances of the same class. Similarly, part-of-speech tagging helps disambiguate word senses and may identify common and uninteresting usages from rare and meaningfully distinct occurrences.

Since statistical methods can be especially efficient, we expect to use them in the initial step of a divide-and-conquer strategy. A rough resolution of a large corpus into topics and subtopics will act as a starting point for finer-grained, more resource intensive, analyses that address issues such as inferring the stance of a particular passage.

Focussed Linguistic Analysis

A persistent difficulty in the design of natural language processing systems is the simultaneous need for coverage and careful analysis. This has made it difficult to reap the expected benefits from the advances in syntactic and semantic theory and parsing technology in the past. While deep analysis is achievable with these techniques it always depends on having available a detailed, hand-crafted, linguistic description with broad coverage—an eternally elusive object.

In contrast, we suggest novel techniques which can dependably detect certain restricted aspects of content. By focussing language analysis, queries can be answered without the use of fragile techniques needed for exhaustive text analysis. Instead of attempting to convert language wholesale into pure information, only directed search is attempted.

As one example, language structure such as syntactic constructions, lexical selection, and even choice of typeface yield evidence about the information in text. When the author's intent is visually or linguistically signalled to the reader, comprehension increases. By the same token, if the author's emphasis can be crudely deduced for use in a program, topic selection improves, even when operating with a small data sample. This stands in contrast to exhaustive inference techniques. It also contrasts with standard IR techniques for extracting information that succeed through (and depend on) large-scale cross-instance comparison.

In addition to functioning as a method for locating significant topics where comparison with previously established categories is not possible, techniques such as the one described for exploiting the focus of a document may also serve to improve the term weighting algorithms which many information search and sorting algorithms depend on. It has long been recognized that the *location* of a word in a text is significant for its role in classification [10]. Typically, the characterization of locations successfully employed has been *proximity* information (i.e. a term is considered more important if it occurs multiple times in close proximity). We recommend that weighting of a term be sensitive to the enriched notion of location, which permits a term to be assigned additional weight when it occurs in the context of an emphatic cue. In short, this approach can also be used to augment traditional information extraction techniques, and simple phrase-spotting and relevance search [2].

Once topicality and significance spotting have identified candidate passages simple syntactic and semantic models can be employed to determine relations between terms. Hearst ([8]) provides one example of this type of technique. Alternatively, if the reduction in the number of sentences left for consideration is substantial enough, detailed syntactic analysis and semantics interpretation into logical representations may even be feasible (see [11] for relevant improvements to the efficiency of constraint solution in unification grammars, and [6] for adaptation of grammar-based interpretation techniques for use in IR systems).

Conclusion

While exhaustive inference is a fascinating goal, the successful location of knowledge across large and diverse corpora represents a more immediate challenge for text-based intelligent systems. Information theater takes advantage of the bandwidth between the presentation and the user, and the complementary strengths of statistical and fine-grain analytical approaches in an attempt to build such intelligent systems.

References

[1] S. K. Card. Information workspaces. In *International Symposium on Next Generation Human Interface Technologies*, 1989.

[2] F. Chen, D. Cutting, P.-K. Halvorsen, J. Kupiec, J. Pedersen, and M. Withgott. Determining significance in spoken discourse and text. Xerox Internal Memo, 1989.

[3] D. Cutting, Jan Pedersen, and Per-Kristian Halvorsen. A browser for full-text retrieval. Manuscript, Xerox Palo Alto Research Center, November 1989.

[4] Jens Erik Fenstad, Per-Kristian Halvorsen, Tore Langholm, and Johan van Benthem. *Situations, Language and Logic*, volume 34 of *Studies in Linguistics and Philosophy*. D. Reidel, Dordrecht, 1987.

[5] T. Fujisaki, F. Jellinek, J. Cocke, E. Black, and T. Nishino. A probabalistic method for sentence disambiguation. In *Proceedings of the International Workshop on Parsing Technologies*, pages 85–94, Pittsburgh, August 1989.

[6] Per-Kristian Halvorsen. Content analysis. Xerox Internal Memo.

[7] Per-Kristian Halvorsen and Ronald M. Kaplan. Projections and semantic description in lexical-functional grammar. In *Proceedings from the International Conference on Fifth Generation Computer Systems,FGCS-88*, Tokyo, November 1988.

[8] Marti A. Hearst. Applying lexical conglomerations to restricted text interpretation. Submitted to AAAI Symposium on Text-Based Intelligent Systems, Stanford, 1990, 1989.

[9] J. Kupiec. Augmenting a hidden markov model for phrase dependent word tagging. In *DARPA Speech and Natural Language Workshop*, October 1989.

[10] H. P. Luhn. The automatic creation of literature abstracts. *IBM journal of research and development*, 2(4):314–319, 1958.

[11] John T. Maxwell, III and Ronald M. Kaplan. An overview of disjunctive constraint satisfaction. In *Proceedings of the International Workshop on Parsing Technologies*, pages 18–27, Pittsburg, August 1989.

[12] Steve Putz. System 33 database overview. Xerox Internal Memo, May 1989.

[13] J. W. Tukey. Internal working papers 1–25. 1989.