Intelligent Text Retrieval

Judith P. Dick and Graeme Hirst
University of Toronto,
Toronto, Ont., CANADA M5S 1A1.
dick@cs.toronto.edu; gh@cs.toronto.edu

1. Introduction

We contend that in order to achieve intelligent text retrieval it is essential to move away from keyword indexes and toward more powerful, detailed representations of text based on linguistic concepts. The retrieval of case law reports has always presented a particularly difficult problem. The legal researcher's goal in searching case law is to determine whether or not there is authority for a point of view. The researcher needs to be able to make associations among selected legal concepts and to navigate among legal concepts with their related facts in order to investigate the issues in a given legal problem. Keyword-Boolean systems make it difficult to satisfy those needs. Indexes to case reports are exceptionally good. We have both topical and factual access. However, we have known the limitations of the approach for a very long time. In 1897, Oliver Wendell Holmes wrote,

> There is a story of a Vermont justice of the peace before whom a suit was brought by one farmer against another for breaking a churn. The justice took time to consider, and then said that he had looked through the statutes and could find nothing about churns, and gave judgment for the defendant. The same state of mind is shown in all our common digests and textbooks. Applications of rudimentary rules of contract or tort are tucked away under the head of Railroads or Telegraphs or go to swell treatises on historical subdivisions, such as Shipping or Equity, or are gathered under the arbitrary title which is thought likely to appeal to the practical mind, such as Mercantile law. (Holmes 1897, p. 59.)

Textual analysis in this domain is especially challenging since each case is unique. Patterns of literary similarity are not common. The reasoning is diffuse, dense, and original. Although reasons for judgement are formally written, the vocabulary of the law is derived from everyday language. Distinguishing the technical meanings of common words is exacting.

Moreover, even though cases are indexed manually with the care appropriate to the subject matter, we find that the character of the most significant element, the argumentation, has been obscured. In order to improve the situation, it is desirable to construct a representation expressive of the conceptual nature of legal decisions.

The key to finding the 'middle-ground' between IR and NLP is, in our opinion, a matter of developing a viable representation for the analysis of a quantity of text. The representation must be unambiguous, but coarse enough to highlight the informational content of the text rather than focus on the literary expression. The representation contained in our recent work, Dick 1991, is just such a representation. The encoding was done manually because the focus of the research was on the retrieval target rather than the process. The aim was to design an expressive text representation suitable for showing the efficacy of conceptual retrieval.

The knowledge base consists of representations of four contract cases. Four pages of printed text were transcribed as sixteen pages of knowledge representation. In addition there is a lexicon of legal concepts, definitions of which, derived from established authorities, have similarly been transcribed, constituting another thirteen pages.

Two of the cases are very simple; two contain complex arguments. The cases come from different time periods and different jurisdictions and so provide a variation of conceptual analysis and linguistic expression. Among the expressions included in the cases are the formidable concepts of 'intention to contract' and the 'foreseeability of consequences' at the instigation of a breach.

An example of text, an excerpt from a simple case, *Stamper* v. *Temple*, is shown in Fig. 1. In it, Turley, J. hedges before stating his opinion. We do not represent the full expression of his doubt, for example, 'constrained', 'to believe', 'what is called. . . '. We do show that the promise is, in his opinion, not an offer, and so concentrate on the informational content of the report to the exclusion of the rhetorical style. The representation of the promise is not included in the excerpt from the representation, which follows the text of the case.

2. The Representational Notation

We used John Sowa's conceptual graphs (cgs) because a fully developed notation is available now. Cgs have a mnemonic aspect and are exceptionally easy for the uninitiated to read. Their use makes bridging the gap between IR and AI audiences much easier. Their expressiveness is attractive for language analysis since it is possible to make some limitations on quantifier scoping, to designate unambiguous coreferents, and to show various kinds of sets with diverse symbols. Furthermore, there is an established user community and we believe that employing cgs brings us closer to the construction of an interpreter because of current software development.

Using cgs did not solve all our representational problems. Although Sowa has provided a catalog of defined conceptual relations — some primitive and some complex — they were not of even quality and did not satisfy the need. Some conceptual relations were simplistic descriptions of complex relations, for example, 'cause' (CAUS) and 'possession' (POSS).

> **cause.** (CAUS) "links [STATE: *x] to [STATE: *y], where *x has a cause *y. Example: *If you are wet, it is raining. [STATE: [PERSON: You]←(EXPR)←[WET]] →(CAUS)→[STATE: [RAIN]]."* (Sowa 1984, 415-416)

> **possession.** (POSS) "links an [ANIMATE] to an [ENTITY], which is possessed by the animate being. Example: *Niurka's watch stopped. [PERSON: Niurka]→(POSS)→WATCH]←(OBJ)←[STOP]"* (Sowa 1984, 418)

Furthermore, some complex ideas are presented as relations and not defined. As a part of a frame for the concept [DEMONSTRATE], Sowa included a graph which apparently says that the purpose (PURP) of the act of demonstrating was a set of demands (Sowa 1984, 262). However, the conceptual relation (PURP) is not named, defined or discussed. Its use appears to be an attempt to deal with the concept of 'intention' which was of great interest in the cases included in our work. We defined additional, supplementary relations. The primitive relations, commonly prepositions, were occasionally used in what appeared to be inconsistent ways. We simple resolved the resulting ambiguities as easily as possible for our own application.

The difficulty of controlling multiple-level embedded clauses, was not entirely overcome. Some devices were contrived to accomplish the representation of long, troublesome sentences. For example, additional subtypes were defined to add structure. In the excerpt below, you can see that we have defined a subtype of the concept [SITUATION], [HYPO], to set apart hypothetical situations described in legal argument from factual situations in the case. Also, [TERMS] was defined as a subtype of [PROPOSITION] which was intended to group graphs. [TERMS] encompasses the contents of an agreement since the meaning of 'contains' (CONT), as defined, was limited to physical contents. The punctuation for contexts, consisting of pairs of dashes and periods for outer contexts and commas and periods for embedded ones, was rather more complicated to handle than was anticipated. Looking at examples from Sowa again raised questions about consistency. We simply adapted the notation to our needs with as much integrity as possible.

Some linguistic cases were included in the catalog as conceptual relations, but they were not adequate to the task of text analysis. They were traditional ones, 'agent', 'patient', 'instrument', and so on. We replaced the Sowa cases with Harold Somers's grid of twenty-eight cases. The grid was designed to resolve some of the worst problem with case — dual roles and the proliferation of cases for handling exceptions to the use of the traditional cases. Application of Somers's cases was successful in that it was not necessary to add any other cases in order to complete the work; and some of the cases were very clearly on the mark. For example, the active source case (ACTS) expressing agency as an initiating quality found in animate and inanimate entities and with or without willfulness (in animate entities), worked very well. This was the case commonly used to represent subject noun phrases, but a number of alternatives were available. In the same row entitled 'source' were a number of other choices for the expression of initiation or agency. One of the most salutary is active local, (ACTL), which neatly conveys the meaning of co-agency, a common representational problem.

Aronstad rules with the Blue Dragons.
[ARONSTAD]←(ACTS)←[RULE]–
(ACTL)→[BLUE_DRAGON: {*}]

In this example, the Blue Dragons are co-agentive with the active source, the agent, Aronstad.

Some other cases, for example, the cases in the objective row were not so clearly defined. They had to do with the expression of undergoing a process. In general they were more passive than the 'active' cases. It was difficult to be clear about the appropriate use of two of them in particular, the local, (OBJL), and the goal (OBJG) cases. Somers description was brief and examples were not included. Moreover, they did not satisfy the need to represent the grammatical object, alone or in combination. In fact, many objects were ultimately assigned to cases along other parameters in the grid. One missed the expression of both the ideas of 'object' and of 'theme' although 'recipient' and 'benefactor' functions were included. Some cases were not suitable for the application for which they were intended. For example, another source case, ambient source (AMBS), was proposed as 'a reason for something', but we have used it sparingly as a designator of causes involving agency. Of course, its use does not solve the causation problem. Throughout, Somers left much to the interpretation of the user and repeatedly stated that the analysis must be adapted to the requirements of the domain.

In spite of the inadequacies, given the framework of the FOL-based Sowa notation, and the valency-based case system of Somers, it is clear that a verb-centered approach to text analysis resulted in a powerful representation of informational content for the purpose of retrieval.

3. Retrieval

We were able to show how retrieval could be accomplished using the representation. Questions were derived from later law cases that followed the contract cases in our knowledge base. The questions were written in the same notation as the knowledge base cases. A frame matching algorithm, LOG (Mezitis 1988), was hypothetically adapted to the task at hand and detailed descriptions of the frame matching in a walk through the proposed searches were written.

LOG produces all the lexical alternatives for an input concept that the system knows. It can handle idioms and can bind internal variables. It also makes use of partial matches. Each concept in an input pattern is treated with equal importance; each is a candidate for a match.

The matching process is done with smart marker passing. The spread of activation is constrained by having some nodes perform as if magnetized. 'Magnets' are the highest generic nodes matched in the initial search attempt. They direct the search to the 'most-likely-to-succeed' paths. This advantage made it possible not only to constrain the search process from making silly hits, but to locate near synonyms as well, so long as they are found close together. However, it does require that the hierarchical organization be meticulous.

Furthermore, we projected enhancing LOG, as LOG+, to include partonomic relations, and to add a little flexibility to adapted searches following failed matches of some types. And we planned a way to negotiate some contextual matches not in LOG's repertoire, in order to take full advantage of those embedded contexts in the representation.

4. Conclusion

We were able to demonstrate how both simple and complicated questions could be matched. Answers included legal concepts, typical fact situations related to specified concepts, definitions of concepts from both the lexicon and the case representations, and best of all, the retrieval of previously unnamed concepts. All of the above were possible in combination. Partial matches were similarly reported. A detailed analysis of the types of matches required for case law retrieval had been prepared. We were able to accomplish, to some extent, all the types of conceptual matches we had comtemplated needing. The initial retrieval example involved a concept for which we had not found index entries in the standard sources. The conceptual representation of the relevant cases, found in legal treatises, made it possible to retrieve those cases using a description of the idea or an associated fact situation. Something that could not be done without a conceptual representation.

The ability to perform inferences, which we were able to demonstrate through the use of the matcher, showed that retrieval capability could be tremendously increased. It was no longer necessary to name exactly the facts or the topic in the way that Mr. Justice Holmes described in the quotation above. Instead, through the use of a caseframe representation, conceptual retrieval of even complex and abstract concepts was possible.

Stamper v. *Temple* (1845) 6 Humph. 113 (Tennessee)

TURLEY, J.: "We are constrained to believe that what is called an offered reward of $200. was nothing but a strong expression of his feelings of anxiety for the arrest of those who had so severely injured him, and this greatly increased by the distracted state of his own mind, and that of his family; as we frequently hear persons exclaim, 'Oh, I would give a thousand dollars if such an event were to happen or vice versa' ".

```
(JD)→[[PROMISE-n: #S1]–
        (~EQUIV)→[OFFER: #S1]
        (CHRC)→[[PHRASE: "EXPRESSION OF STRONG FEELING"]–
                    (EQUIV)→[EXPRESSION: #S1]–
                                (OBJL)→[[[FEELING: #S1]→(ATTR)→[STRONG: #S1]]
                                  or [ANXIETY: #S1]]
        (AMBS)→[ANXIOUS_FOR: #S1][STATE_OF_MIND: #S1]
    [HYPO: [PROMISE-n: #S2]→(CONT)→[TERMS:
          if [HAPPEN: #S1]–
              (ACTS)→[EVENT: *a]
          then [GIVE: #S2]–
                (ACTS)→[PROMISOR: *m]
                (DATPOSSL)→[REWARD: #S2]→(MEAS)→[MONEY: @$1,000]].
      [PROMISE-n: #S3]→(CONT)→[TERM:
       if ~[[HAPPEN: #S2]–
            (ACTS)→[EVENT: *b]]
        then [GIVE: #S3]–
              (ACTS)→[PROMISOR: *n]
              (DATPOSSL)→[REWARD: #S2]–
                          (MEAS)→[MONEY: @$1,000].]]];end of hypo
```

Fig. 1. A case excerpt and a part of its representation.

References

Dick, Judith P (1991). *A conceptual case-relation representation of text for intelligent retrieval.* PhD dissertation, forthcoming, August 1991. Toronto: University of Toronto, 1991.

Holmes, Oliver Wendell (1897). "The path of the law," In MacGuigan, Mark Rudolph (editor) (1966). *Jurisprudence: Readings and cases.* Toronto: University of Toronto, 1966, 48–62.

Mézitis, Mara Anita (1988). *Generating lexical options by matching in a knowledge base.* Technical Report CSRI-217. Computer Systems Research Institute, University of Toronto, September, 1988.

Somers, Harold L (1987). *Valency and case in computational linguistics.* Edinburgh: Edinburgh University Press, 1987.

Sowa, John F (1984). *Conceptual structures: Information processing in mind and machine.* Reading, MA.: Addison-Wesley, 1984.