INFORMATION RETRIEVAL IN THE SERVICE OF GENERATING NARRATIVE EXPLANATION

What we Want from GALLURA

Ephraim Nissan¹ and Yaakov HaCohen-Kerner²

²Department of Computer Science, Jerusalem College of Technology (Machon Lev), P.O.B. 16031, Jerusalem, Israel

Keywords: Information extraction, Explanation generation, Story generation.

Abstract: Information retrieval (IR) and, all the more so, knowledge discovery (KD), do not exist in isolation: it is necessary to consider the architectural context in which they are invoked in order to fulfill given kinds of tasks. This paper discusses a retrieval-intensive context of use, whose intended output is the generation of narrative explanations in a non-bona-fide, entertainment mode subject to heavy intertextuality and strictly constrained by culture-bound poetic conventions. The GALLURA project, now in the design phase, has a multiagent architecture whose modules thoroughly require IR in order to solve specialist subtasks. By their very nature, such subtasks are best subserved by efficient IR as well as mining capabilities within large textual corpora, or networks of signifiers and lexical concepts, as well as databases of narrative themes, motifs and tale types. The state of the art in AI, NLP, story-generation, computational humour, along with IR and KD, as well as the lessons of the DARSHAN project in a domain closely related to GALLURA’s, make the latter’s goals feasible in principle.

1 CONCEPTUAL & TECHNICAL BACKGROUND

In the history of full-text IR, tools for retrieval from very large historical corpora in Hebrew and Aramaic were prominent, with the RESPONSA project (see e.g. Choueka, 1989a, 1989b; Choueka et al. 1971, 1987). Before the rise of Web search engines, RESPONSA tools were the ones which achieved the more far-reaching effects on society, because how they empowered the retrieval of legal precedents in rabbinic jurisprudence, thus affecting especially legal practice of family law in Israel (as for family law, in the Ottoman successor states, the usual jurisdiction is the courts of the various religious communities).

Religious cultures, as being the “consumers” of religious texts, were, in a sense, the customers of a considerable portion of early projects in IR: apart from RESPONSA, whose corpora comprise the Jewish texts from the sacred sphere through the ages, this was also the case of Padre Busa’s Index Thomisticus in Milan, and of the humanities computing at the Abbey of Maredsous, in Belgium.

Exegesis (such as biblical interpretations) and homiletics involve layers of texts, where a secondary text refers to and either just quotes, or discusses, some locus in the primary text; or then (as in the Jewish aggadic midrash) expands on a biblical narratives, filling the gaps where the primary text is silent. Collections of aggadic midrash from late antiquity (e.g., the Midrash Rabbah) or the Middle Ages (e.g., Yalqut Shim‘oni) are a digest of a multitude of homilies on biblical fragments of texts, developing several often alternative ideas and subnarratives. Cf. Hirshman (2006), Braude (1982), Fishbane (1993), Hartman and Budick (1986).

* HyperJoseph is a hypertextual tool on the story of Joseph in Genesis, with the secondary texts elaborating on it (Nissan and Weiss, 1994).

* DARSHAN is a tool that invents homilies in Hebrew (HaCohen-Kerner et al. 2007). Retrieval in DARSHAN is intensive, and so is the use of networks of lexical concepts.

DARSHAN generates ranked sets of either one-sentence or one-paragraph homilies. While producing its output, DARSHAN is able to quote...
from Scripture, to search for an occurrence elsewhere in the textual canon, to replace words or letters, to resort to puns, to interpret a word as an acronym, and so forth. Use is made of patterns which consist of canned text with places where to plug in strings obtained through IR and manipulation. The user supplies as input a biblical verse, or a sentence, or a set of words, and also specifies which devices should be applied. Filters applied to the candidate output are alert, e.g., to positive vs. negative connotation.

The quality of an individual output homily is assessed as a sum of weighted factors, including: length (as an indicator of complicacy); the percentage of relevant words in the homily, out of the total of words in the homily how many sentences there are; how complex it was to insert every motif into the homily generated; how many motifs were actualized in the output homily being evaluated; how many transformations were carried out; how many words were replaced in the homily.

Having mentioned acronyms, consider that HaCohen-Kerner et al. (2010b) discussed an abbreviation disambiguation system for rabbinic texts in Hebrew or Aramaic. Cf. Stock and Strapparava (2005) on the HAHA project, whose purpose is the humorous interpretation of acronyms. As to connotations, Strapparava and Valitutti (2004) described an affective extension of WordNet.

2 FUNCTIONS IN GALLURA

The GALLURA project seeks to develop software that would interpret in Hebrew names by folk-etymology, but in the context of a generated narrative (aetiological tales, usually brief or even very brief). The most closely studied model is a large textual corpus of playfully creative writing that embodies midrashic literary devices, by explaining fancifully place-names of names for animal kinds.

The GALLURA project, now in the design phase, requires, among the other things, capabilities of story-generation, and of generating a playful explanation. By themselves, these two tasks draw upon three areas in AI:

- explanation synthesis (for which, see e.g. Schank, 1986, 1994; Walton, 2004),
- story-generation (see e.g. Liu and Singh, 2002; Löhneker et al., 2005; and a long survey in Nissan, 2011a: Ch. 5), and
- computational humour (see e.g. Stock et al., 2002; Ritchie, 2004; Waller et al., 2009). Humour studies are interdisciplinary.

Moreover, GALLURA needs skills from computational linguistics, including some that thus far were modelled by linguistics, but not computationally:

- folk-etymology (see e.g. Kirwin, 1985; Coates, 1994; Baldinger, 1973; Zuckermann, 2006), and
- phono-semantic matching (PSM), a discussion of which is found in Zuckermann (2000, 2006).

For example, one of several PSM rules as occurring in neologisation by adapting the meaning of a pre-existent word in the target-language (TL). Another rule of camouflaged borrowing (ibid.) is:

```
SL y ‘b’ → TL(+PSM) x ‘b’ ↵ TL x ‘a’
```

\( x \) is phonetically similar to \( y \); \( a \) is similar to \( b \)

That is to say, the PSM introduced a new sense: this was a PSM produced by shifting the meaning of a pre-existent word in the target-language (TL).

```
SL y ‘b’ → TL(+PSM) \{x\}+\{z\} ‘b’ ↵ TL \{x\} ‘a’, \{z\}
```

\( x \) is a lexical morpheme (e.g. root) that is phonetically similar to \( y \);
\( z \) is a grammatical morpheme (e.g. noun-pattern);
\( \{x\}+\{z\} \) is one word; \( a \) is similar to \( b \)

GALLURA should also have quality evaluation capabilities, e.g., evaluating a story generated (Peinado and Gervás, 2006), or evaluating morality within a story (Reeves, 1991). We also need to resort to computational argumentation: some such current research into argumentation in computer science looks into legal narratives (Bex, 2011).

Explanation as sought in GALLURA need not necessarily be realistic: it is non-bona-fide (like in humour), and must conform to a set of conventions, of which realism is just a particular case (cf. Nissan, 2008). There are constraints on style: the output text generated conforms to the early rabbinic linguistic stratum and style (thus emulating the aggadic midrash), with constraints on which lexical items or morphological forms can be selected.

Rabbinic stylomes are the subject of current IR research, including in the CUISINE text classifier. So are the identification of rabbinic citations, and chronological classification based on them. In fact, HaCohen-Kerner et al. (2010a) discussed stylistic feature sets for classification in CUISINE. Automated identification of citations from rabbinic texts has been researched (HaCohen-Kerner et al., 2010c). Automated classification of rabbinic
responsa by period based on what they cite or are cited by, was attempted successfully: HaCohen-Kerner and Mughaz (2010) defined and effectively applied “various kinds of ‘iron-clad’, heuristic and greedy constraints defining the birth and death years of an author based on citations referring to him or mentioned by him.”

3 THE MULTIAGENT ARCHITECTURE

Several capabilities are required of GALLURA, and many of them require retrieval. Fig. 1 shows Coalition1 of agents, i.e. agents that often interact among themselves. The control sequence is opportunistic, according to the needs of the various agents while they tackle a (sub) problem during a particular run. They broadcast their need for help to the other agents, and contract out the task. Some agents however interact in a privileged manner with one or more other agents, as they form a “coalition”.

Both the syntax agent and the stylemic agent have to emulate early rabbinic language, but the pool of stylemes and more abstract modes comprising stylemes need actually be wider. Fig. 2 shows the interplay of other coalitions of agents. In the Lexicon, expected associations or behaviour are triggered through demons, procedural code activated upon access to individual lexical entries.

Coalition5 comprises an Encyclopedic agent, and a Commonsense agent. The latter comprises two modules: Concept-centred commonsense, and Situational commonsense. Both the Emplotment agent, and the Tex-generation agent closely interact with the Argumentation agent.

4 A SIMPLE EXAMPLE: AQUA & GENESIS 1:9

It is usually proper nouns that are playfully etymologised in the modern, archaising Hebrew narrative corpus which is the main model for GALLURA, and whose own model is the already mentioned early rabbinic genre of the aggadic midrash. Nevertheless, sometimes common nouns are folk-etymologised as well, and most often these are non-Hebrew words.

Here is a concise example. The input is Latin aqua ‘water’. In the model corpus, there is this item:

Ma ra‘û Bnei Rómi, še-hém qorín et ha-máyim ‘aqwa (aqua)? Le-fí še-katúv: “yiqqawú ha-máyim”.

Here is a translation of this Hebrew text:

Why [literally: what did they see], the Romans [lit.: The Sons of Rome], that they call water aqua? Because [lit.: to mouth of] it is written [in Scripture]: “Let the water be gathered”.

In fact, the intertextual reference is to Genesis 1:9. The verbal form yiqqawú (passive future, 3rd
5 A COMPLEX EXAMPLE: BABEL TO LAOS

The following would be a much more difficult example for GALLURA to replicate, and both retrieval and manipulation would be intensive and laborious. In the model corpus we use, place-names around the world are explained by both playful etymology, and fantasy history narratives. It is often the case that a story is told about one of the human groups leaving the Tower of Babel. The Generation of the Division (Dór ha-Pallagá) or the Ones Leaving the Tower (Yots'éi ha-Migdal) would be often resorted to in GALLURA’s Pool of stock characters. Let us consider a story on Laos.

“Teach us, Sir” (velammedenu Mar, a cliche especially associated with the lost rabbinic Midrash Yelammedenu), “What did the Nations see” (i.e., ‘why’: má rá’á ha-’ummôt), “that they call” (šegorín: a Mishnaic verbal inflection) “one of them Laos” (achát mehén Lá’ós). “I shall answer you immediately!” (Af ani mešikhem mi-yád/ a cliché). A ready pattern of argumentation: “Instead of [lit.: Until] you asking why that nation is called Laos” (‘Ad šé-attém šo’alin lamma otáh ’ummá qruyá Lá’ós), “be asking what did the Sons of Greece see” (hevá šo’alin ma rá’á Bnei Yaván), “that all populations” (še-kál ’okhlosin, itself a Green loanword in Hebrew) “were called in their mouths [i.e., by them] λαος” (niqre’ú be-fihém lá’ós). “Once the Ones Leaving the Tower went out of Babel” (Keván šé-yats’ú Yots’éi ha-Migdál mi-Bavél), “they were tired (le’in) and exhausted on the road” (hayá le’ín u-me’oyafín ba-dárek). Sustained walking is tiresome, and one term for ‘tired’ is related by PSM to Laos. Now, consider that in a crowd (a spawned demon would inform GALLURA), you would expect somebody trying to sell snacks and drinks, unless circumstances exclude this (e.g., if it’s a day of fast, or a famine causes starvation). Such a situational cliché is funny if it does not quite match the situation at hand. The theme of the exodus from Babel, in the model corpus, often has a wise old man advise the crowd, but some other time, some individual takes advantage, being cunning rather than altruistic.

“The more astute among them” (‘Armamiyyín sebahém), “who were traders and vendors of edibles” (še-hayá ba’aléi praqmátya [a typical early rabbinic term] u-mokhréi mezonót), “this way they were speaking to them” (kákha hayá ‘omrim lahém): “Let the legs be strong!” (Techezáqna ha-raɡláyim/). The latter contains a Biblical Hebrew verbal form, the 3rd person plural feminine (as ‘legs’ are feminine in Hebrew), whereas Mishnaic Hebrew discarded that form, using the masculine. As this is a modified quotation, using a Biblical Hebrew morphological (or lexical) form is legitimate for GALLURA. “Let the legs be strong!” (Techezáqna ha-raɡláyim/) is a modification of “Let the hands be strong!” (Techezáqna ha-yadáyim/), the title of a famous labour song by Bialik. Such a temporal flashforward for a story set at the times of the Tower of Babel is a funny transgression (rather than an insipid inconsistency).

“Whatever you shall put under your teeth, you shall find in your legs!” (Má še-tittní táchat šinnekhékim, timtse’ú be-raɡlékhin/). This is a Hebrew adaptation of an Aramaic early rabbinic proverb. “Be chewing” (hevá lo’asín, associated by PSM with Laos), “as for this you were created!” (še-
la-zé notsártēm! This is evocative of le-khakh notsarta, “for that purpose [of studying] thou hast been created”, in Maxim of the Fathers, 2:8).

“Every population” (lit.: population population, ‘okholisin ‘okholisin, “all of them are chewers!” (kullam la’osot!). “As they were hearing them saying so” (Keván še-hayú som’in ‘otám ‘omrim ken, partly a quotation of how the crowd in the Temple used to respond to a given utterance of the High Priest on the Day of Atonement, a day of fast), “their saliva flowed, they paid the price, would take and eat” (zāv hayá virám, notnin mamón, notlin ve-‘okhlin, with typical Mishnaic wording).

Clearly, obtaining from GALLURA output such as this story from our model literary corpus would be as ambitious a goal as it can get. Anything in the middle would be nice to achieve. See Nissan’s (2011b) 150-page discussion of playful narrative explanations.

6 WHENCE AND WITHER? VON IN GERMAN ONOMASTICS, AND ELDAD & MEDAD

It is important to realise that knowledge discovery or information extraction as involved in accessing the historical textual canon as well as onotologies and representations of commonsense or encyclopedic knowledge, can be easier in one direction, while very difficult in the other. We exemplify this with an item from our model corpus. Edom (medieval for ‘Europeans’) in the Land of Ashkenaz (medieval for ‘Germany’) the text relates, for many generations were eager to insert fon (i.e., von) before their family names, as it would signal their patrician ancestry.

“What did cause that? The episode of Eldad and Medad caused that”, at Numbers 11:26–29. Moses appointed seventy elders, but those two did not come, and prophesised nevertheless. Joshua tells Moses to put them under arrest, but Moses retorts: “Are you jealous on my behalf? If only” all the people were prophets.

U-mi yiṭṭēn (lit.: “And who would give”) was rendered, in the canonical Jewish Aramaic translation (the Targum by Onqelos) as: Ra’ena fon, i.e., “I wish fon”, where fon (a grammaticalised denominational conjunction) means any of ‘face’, ‘turn’, ‘that would’, or ‘lest’. If you were reading Onqelos, you may happen to notice this locus serendipitously. But had you begun with eagerness for ennoblement, it would be very difficult to devise an appropriate search that would retrieve a biblical “I want fon”.

7 CONCLUDING REMARKS

GALLURA is an ambitious project, now in the design phase, requiring the interplay of various agents or coalitions of agents specialised per domain of expertise. Several of these agents have retrieval-intensive requirements. GALLURA has to devise playful etymologies with a backup story to go with. It builds upon the experience and part of the architectural features of DARSHAN — especially how the pool of devices is organised, and the approach to retrieval, which is mostly from the same textual corpora. GALLURA is much more difficult to achieve, but at the stage reached by a number of domains within AI, NLP, IR, and KD, it is in principle feasible. Any progress on any part of the architecture would by itself be a valuable achievement. A global advantage already at present, in this project, is that thanks to manual analysis of many items in the creative writing corpus which is our main model, it is possible to model algorithmically all devices required.

REFERENCES


491


Hartman, G. H., Budick, S., eds. 1986. Midrash and Literature, Yale University Press, New Haven, CT.


