An interactive natural-language genealogy quiz engine

Genealogical information is data-rich: it includes many items of low-level information such as dates, names, locations, family relationships, and documentary references. Typical repositories of such information include GEDCOM files; they often contain information on hundreds, or even thousands, of individuals and families. Becoming familiar with the contents of such data files can be a daunting task. Last year's workshop included presentations on how GEDCOM information can be accessed and analyzed via various techniques including data visualization and speech. In these scenarios the initiative belongs to the user of the system. This presentation focuses on how such data can be used to drive an interactive natural-language game where the system has the initiative.

While the commercial viability of such an engine has yet to be demonstrated, it is not difficult to motivate possible applications. Three factors combine to suggest that such an engine might be helpful. First, genealogy work is not often viewed as a fun, entertaining enterprise, and consequently almost no technology has been developed to date with this perspective. Second, it is often difficult to gain a global impression from the low-level facts presented in a typical GEDCOM file, and simply browsing the data in one of the commonly available interfaces (PAF, for example) does not give this perspective. Third, the demographics of family history research has (at least presumably) involved primarily older persons. Computerization of such work is progressing quickly, and older persons are sometimes less adept at the use of computers. Conversely, children and youth are less commonly involved with genealogical research but highly computer-literate. Thus this engine is meant to provide an entertaining way for researchers, novice or experienced, to gain a higher-level appreciation of the data in a GEDCOM repository, and to test and review such knowledge.

The engine combines several techniques from the fields of natural language processing and human/computer interaction, and integrates various components designed for similar tasks. A GEDCOM file is supplied to the engine, which then parses out the file's information and stores it as a PROLOG database. This propositional logic format is important since PROLOG supports forward inferencing and other goal-directed reasoning techniques. Once the information has been stored, a set of pre-specified inferential relationships is automatically generated by the system. Questions might involve specific data items about an individual (e.g. Where was your paternal grandfather born?) or might be of a very global nature (e.g. Name two of your ancestors who immigrated to America.).

Additionally, at run-time a minimal amount of information about the user is also supplied (via user-selectable menu or perhaps via dialogue (textual or speech). The primary purpose for this information is twofold: to situate the user with respect to the other information in the file, and to ascertain the level of expertise of the user with respect to the data in question (e.g. minimal=very little, average, expert=very knowledgeable). This helps the system set an appropriate level of specificity and difficulty for the questions it will ask.

Once the system has been initialized and the data compiled, the engine enters into an interactive, goal-directed dialogue with the user. The system presents to the user a series of family history questions for which one or more alternative answers have been determined from the fact base. The dialogue structure of the engine follows state-of-the-art techniques including a dialogue move engine developed for multi-participant goal-directed discourse. Questions are generated from propositional content of the knowledge base via a phrase-structure grammar designed specifically for the task. The system gauges the correctness of the user’s response(s) for each question and responds accordingly.

Input/output modalities of the system are still being finalized. At the very least, the system is able to communicate with the user via a keyboard interface. It is assumed that interactions will also be possible via speech input/output, using for example, the OGI speech application toolkit. Interactions may also include presentation of multimedia data such as pictures or sound or video clips if such data is provided in the GEDCOM file.

The contributions of this work include novel integration of NLP techniques with GEDCOM data, the provision of complex discourse structure for discussing family history facts of both specific and global nature, and the development of a new form of entertainment engine that could be useful in helping create sustained interest in genealogical information.