WikiNetTK – A Tool Kit for Embedding World Knowledge in NLP Applications

Alex Judea, Vivi Nastase, Michael Strube
Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany
{alexander.judea, vivi.nastase, michael.strube}@h-its.org

Abstract

WikiNetTK is a Java-based open-source toolkit for facilitating the interaction with and the embedding of world knowledge in NLP applications. For user interaction we provide a visualization component, consisting of graphical and textual browsing tools. This allows the user to inspect the knowledge base to which WikiNetTK is applied. The application-oriented part of the toolkit provides various functionalities: access to various types of information in the knowledge base as well as methods for computing association paths and relatedness measures. The system is applied to a large-scale multilingual concept network obtained by extracting and combining various sources of information from Wikipedia.

1 Introduction

Since the early 1990s the quest for large scale machine-readable knowledge repositories has become more and more intense. Cyc (Lenat and Guha, 1990) relies on experts to add to its knowledge base, MindPixel and Open Mind Common Sense (Singh, 2002) opened the contributors base to everyone using the Internet as a collaborative platform. Research in the 2000s has focused much on Wikipedia and extracting the knowledge it provides for human consumption into machine readable format. DBpedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007) are two such examples. With the extraction of very large knowledge repositories it becomes crucial to provide tools for the user that would allow her to explore and use the information contained therein.

We introduce WikiNetTK (WNTK), a tool for visualizing and exploiting world knowledge extracted from Wikipedia. It is composed of three main parts: (i) an API that serves as an interface to the knowledge base used; it is currently applied to WikiNet, a concept network extracted from Wikipedia (with minor modifications this can be adapted for other, similar, knowledge bases); (ii) a visualization component, that allows the user to inspect the knowledge encoded in the resource; (iii) functionalities for computing association paths between concepts and computing semantic relatedness. WNTK also provides a command line interpreter for users who wish to work outside the Java environment; the commands implemented so far allow the user to retrieve and output concepts, paths between concepts, access the relatedness metrics and use the visualization component to directly visualize concepts or paths.

WikiNetTK[1] is an open-source Java-based system. For data management it uses BerkeleyDB[2] and for visualization prefuse[3].

2 Data

WikiNetTK is applied to WikiNet, a repository of world knowledge extracted from Wikipedia (Nastase et al., 2010). It is derived from the category and article network, disambiguation, redirect, cross-language, infobox and textual content of Wikipedia pages. It is organized as a concept network – it separates concepts and their lexicalizations, and contains relations between concepts – in a manner similar to WordNet. Concepts have lexicalizations in numerous languages. With WikiNet's 3.7 Million concepts and 40 Million relations (instantiating 656 relation types), efficiency in data management becomes an issue. Manual analysis of the data is also problematic. WikiNetTK addresses both these issues. A fast data management is the basis for the user's computations and for an easy-to-use visualization component.

[1] http://sourceforge.net/projects/wikinettk/

[2] http://oracle.com/technetwork/database/berkeleydb

[3] http://prefuse.org

Proceedings of the IJCNLP 2011 System Demonstrations, pages 1–4,
Chiang Mai, Thailand, November 9, 2011.

3 System components

3.1 API

The API provides the interface between database management and data usage. This separation allows a user to easily customize the database management system (DBMS) to her needs, without an impact on the functionality of the system, as well as change the knowledge base used. WNTK's database mainly contains the following Java objects, reflecting the organization of knowledge:

1. *Concept*. A *Concept* contains a flag indicating if it denotes a named entity, the number of hyponyms it has in the network, its network depth, a definition (the first sentence in the respective Wikipedia article), a set of semantic relations to other concepts, a set of names in different languages, and a unique ID.

2. *RelationType*. Each relation type in the database is substituted with a unique ID. This has the advantage of reducing memory usage and allows for a relation to have various names, the same as for concepts. A *RelationType* provides the name of a relation type (e.g. "IS_A"), along with its frequency.

In the interaction with the database, an ID will be resolved to its corresponding concept, a term (e.g. "book") will be resolved to a set of possible concepts, a concept can be expanded with its related concepts up to a maximum distance, and we can obtain paths between concepts in the network.

To avoid re-doing expensive computations and excessive database accesses, the API provides an extra cache for computed paths and expanded concepts (represented as vectors).

Every functionality of WNTK (e.g. the visualization component) expects an abstract type of the API – which means that the user has to reimplement only a few basic I/O related methods to be able to exchange the entire database management or the data used. The actual WNTK distribution comes with Berkeley DB Java Edition, a fast, cache-based DBMS.

3.2 Visualization

When presented with a large scale machine readable repository of knowledge, manual inspection is desirable, but problematic. WNTK's visualization component is an intuitive and efficient way to examine the underlying network, in our case, WikiNet. The user can choose between a graphical network visualization, a text-based concept and path browser, which we present in Figure 1.

The user can type a term (e.g. "book"), and then choose from a set of possible concepts. Words are ambiguous. In WikiNet, in particular, concept names come from different sources: *canonical names* come from Wikipedia article titles, *aliases* come from the redirect, disambiguation pages and cross-language links. To help the selection of the concept to be visualized, the definition is shown as a tool tip when the cursor hovers above the respective list item. Once a concept is chosen it is displayed according to the visualization style, and the user can continue the exploration by clicking on the relation clusters (in the graphical version) or on the hyperlinks (in the text version).

3.2.1 Graphical Visualization

The selected concept is rendered as a node in the middle of the canvas, surrounded by its relation types. The caption of a relation type node is its respective name and the number of relations its concept has to other concepts with this particular type. For example, if a concept has seven "IS_A" relations to other concepts, the caption of the node will be "IS_A: 7". This kind of aggregation keeps the amount of rendered nodes as low as possible. The user can select which relation type node to expand, and thus explore only the parts she is interested in, leaving the rest aggregated.

Although the rendering system[4] tries to avoid overlapping edges and nodes by re-arranging them, the number of rendered elements can become very high and confusing. Parts of the displayed network can be highlighted by hovering with the cursor over concept or relation type nodes – all the nodes they are directly connected to will change color. An example is presented in the first two screenshots in Figure 1.

3.2.2 Text-based browsing

The text-based browser works with a hyperlink structure, as shown in Figure 1. The upper part of the browser field displays the number of hyponyms, named entity information, the definition, and all names. The list of names is collapsed by default and can be shown if needed. The lower part contains the concept's relations, grouped by its relation types. Every hyperlink can be explored. A history keeps track of the exploration. The text-

[4] *prefuse* is used to render the nodes and edges and to rearrange the nodes.

based browser is a good way to explore many concepts or relations in short order.

3.2.3 Text-based path browsing

In the text-based path browser the user can choose two concepts and a maximum path length. All paths between the selected concepts are then computed and displayed. If the selected concepts are both children of a concept (i.e. a common subsumer), this concept will be displayed in bold face. When the user clicks a hyperlink, the respective concept is shown in the text-based browser. The last screenshot in Figure 1 shows the usage of the path browser.

3.3 Functionalities

The API provides fast access to the knowledge base, including retrieving concepts (through their ID or lexicalizations) and their relations. Apart from these basic operations, WNTK provides methods for retrieving paths between concepts, and compute similarity, which are basic tasks for which lexical/knowledge sources are used in NLP. At this point, the toolkit contains several implementations of semantic relatedness measures, in particular Jiang and Conrath (1997), Lin (1998) and Resnik (1995) which were shown to have highest correlation with human judges on WordNet (Budanitsky and Hirst, 2006) as well as several customized measures. The user can also retrieve association paths between concepts. The set of methods can be extended by the user, and other functionalities can be added as well. We are currently working on integrating a module for text annotation relative to the embedded resource.

4 Command line tool

Our purpose was to provide a tool that facilitates the integration of world knowledge in NLP applications. For the users who do not wish to edit or interact directly with the Java source code, WNTK provides a command line interpreter constituting an intermediary layer between using the visualization component and using the API programatically. Because of increased load time of the database, the API is initialized once. After that, the user can access the information in the knowledge base through the available commands:

1. *gc*. A command to retrieve and output concept and relations information in different ways. The command handles concept IDs and terms of various length in the same way.
2. *visual*. A command to start the visualization component. It can be provided with none, one or two arguments, causing the visualization component to be initialized in different states (starting state, concept visualization, and path visualization).
3. *rel*. A command to compute semantic relatedness between any pair of terms or concept IDs, using any of the implemented relatedness measures.

Each command has a manual page, which can be accessed using *man*.

Acknowledgements

This work has been partially supported by the EC-funded project CoSyne (FP7-ICT-4-24853) and by the Klaus Tschira Foundation.

References

Sören Auer, Christian Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a Web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference,* Busan, Korea, November 11-15, 2007, pages 722–735.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING).*

Douglas B. Lenat and R. V. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Addison-Wesley, Reading, Mass.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* Madison, Wisc., 24–27 July 1998, pages 296–304.

Vivi Nastase, Michael Strube, Cäcilia Zirn, Benjamin Boerschinger, and Anas Eghafari. 2010. WikiNet: A very large-scale multi-lingualc concept network. In *Proceedings of the International Conference on Language Resources and Evaluation* Malta, 19-21 May 2010, page to appear.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montréal, Canada, 20–25 August 1995, volume 1, pages 448–453.

Push Singh. 2002. The Open Mind Common Sense project.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of the 16th World Wide Web Conference,* Banff, Canada, 8–12 May, 2007, pages 697–706.

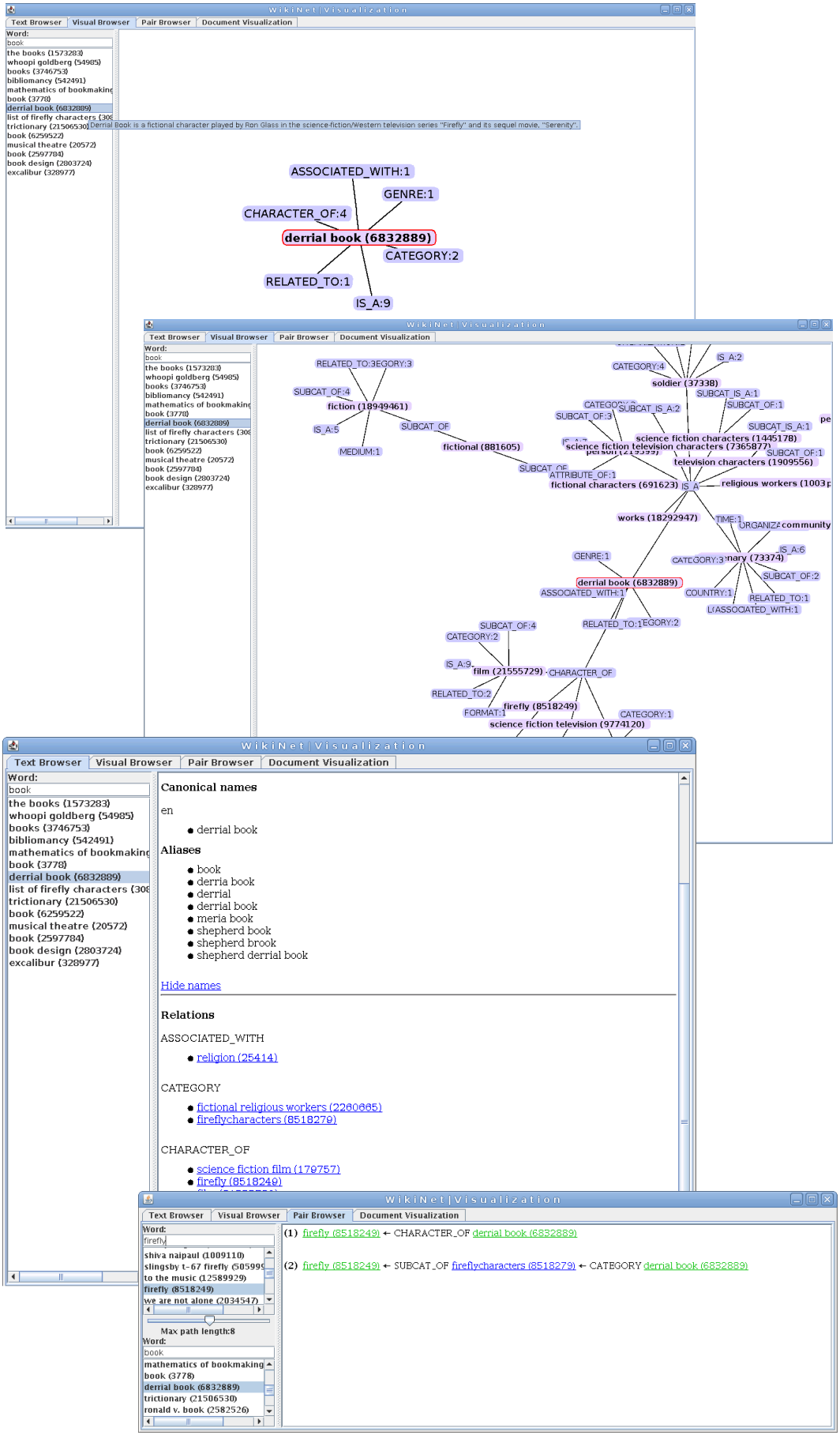

Figure 1: Graph-based, text and path visualization