Kanopy: Analysing the Semantic Network around Document Topics

Ioana Hulpuş[1], Conor Hayes[2], Marcel Karnstedt[1], Derek Greene[3], and Marek Jozwowicz[1]

[1] Digital Enterprise Research Institute, NUI-Galway, Ireland
ioana.hulpus@deri.org, marcel.karnstedt@deri.org, marek.jozwowicz@deri.org
[2] College of Engineering and Informatics, NUI-Galway, Ireland
conor.hayes@nuigalway.ie
[3] School of Computer Science and Informatics, University College Dublin, Ireland
derek.greene@ucd.ie

Abstract. External knowledge bases, both generic and domain specific, available on the Web of Data have the potential of enriching the content of text documents with structured information. We present the Kanopy system that makes explicit use of this potential. Besides the common task of semantic annotation of documents, Kanopy analyses the semantic network that resides in DBpedia around extracted concepts. The system's main novelty lies in the translation of social network analysis measures to semantic networks in order to find suitable topic labels. Moreover, Kanopy extracts advanced knolwedge in the form of subgraphs that capture the relationships between the concepts.

1 Introduction

Recent research has made progress in interlinking text documents with the Web of Data. One of the main benefits of this linkage is that the external knowledge can be used as a complementary source of information, enriching the content of the original documents and revealing semantic relations between them. There exist several popular systems for this task. Zemanta [13] focuses on enriching blog posts by recommending the authors publicly available content that can be added to the post, for example, images or links. OpenCalais [11] aims at annotating text with semantic entities and events that are extracted from it. WikipediaMiner [9] provides links to corresponding pages in Wikipedia [12] and related concepts, while DBpedia Spotlight [5] focuses on linking entities from text to those in DBpedia [6].

However, most of these systems just annotate the text with the disambiguated concepts extracted from external data sources. While this does add value to the user experience, we argue that this leaves most of the potential unexploited. DBpedia and other semi-structured knowledge bases offer a wealth of exploration options available with comparatively low processing costs. The relations between

concepts can be analysed together with the graph structure around them, resulting in the discovery of rich knowledge that is not necessarily obvious from the text itself. Advantages of such an analysis are manifold. Besides concept linking, it can serve to: (i) provide explanations for concept linkage; (ii) enrich texts with a wealth of background knowledge; (iii) provide starting points for further knowledge exploration; (iv) provide insights in the quality/quantity of information the knowledge base contains about the topics discussed in the text reveling possible knowledge gaps.

Our system *Kanopy* demonstrates these advantages. It extracts a list of relevant topics from an input document, where each topic consists of a set of related words. The main objective of Kanopy is to automatically label each topic with a concept that captures its essence. Thus, for each topic Kanopy returns the topics, the top k recommended labels as well as a *topic signature graph*. This graph shows the relations between the topic words and the suggested labels, as well as other strongly related concepts. In order to achieve this, Kanopy tackles a number of difficult problems: keyphrase extraction from text, topic finding, concept linking and disambiguation, graph extraction and topic labelling.

An important approach related to our work is the REX system [1]. Given a pair of entities and a knowledge base (i.e, DBpedia), REX extracts a ranked list of semantic paths that explain the relationship between the two concepts. MING [4] is another related system that, given a group of concepts and a knowledge base, returns the most informative subgraph that contains all the given concepts and the relations between them. These systems are focused on extracting the relations between the seed concepts. Kanopy's main contribution lies in applying graph-based centralities to rank related concepts and extract the ones suitable for labelling the topic. The resulted concept must be central, from a semantic-graph perspective, with respect to all the topic words. In order to measure this *semantic centrality*, we analyse the semantic network that interconnects the concepts behind the words. When combined with a convenient user interface, Kanopy offers knowledge discovery beyond that offered by simple topic labelling.

2 The Kanopy System

In this section, we overview the key stages of the Kanopy processing pipeline, as illustrated in Figure 1.

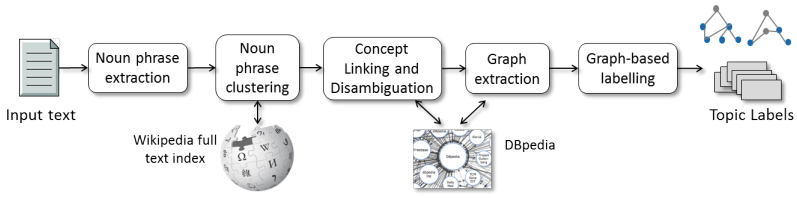

Fig. 1. Diagram illustrating the complete Kanopy processing pipeline.

1. Noun phrase extraction. The main noun phrases are extracted from the raw input using the Stanford CoreNLP library [8]. These phrases are weighted and then filtered according to their TF-IDF score, computed with respect to the Wikipedia full-text corpus.
2. Noun phrase clustering. As in probabilistic topic models, we assume that a text document contains one or more sets of related "meaning-bearing words", where each set corresponds to a different topic. To identify these topics, we cluster the noun phrases obtained from the previous stage using agglomerative hierarchical clustering. Positive Pointwise Mutual Information (PPMI), with respect to the Wikipedia full-text corpus, is used as similarity metric. The applied linkage strategy and cut-off point can be adjusted from the user interface. After this stage, we obtain several clusters of noun phrases.
3. Concept linking & disambiguation. This stage links and, if necessary, disambiguates every noun phrase in each obtained topic to concepts from DBpedia. While many algorithms exists for word-sense disambiguation (WSD), we use the Eigenvalue-based WSD [2]. It is unsupervised, does not need preprocessing and it supports simultaneous disambiguation of a group of related words. It achieved approximately 10% better accuracy at disambiguating topic models than the state-of-art unsupervised WSD algorithms [2]. Its drawback lies in the higher computational complexity, but the parallel implementation inside Kanopy meets the requirements of an online demonstration.
4. Graph extraction. At the beginning of this stage, each remaining noun phrase is linked to exactly one DBpedia concept, which we refer to as *seed concepts*. A semantic network is extracted by activating the concepts and their relations within a distance of two hops from each seed. In the majority of cases, this is sufficient to connect the single concept subgraphs into one larger graph, which provides the input for the next stage [3].
5. Topic labelling. From this extracted topic graph we finally identify the most relevant concepts. We apply the *focused random walk betweenness* and *focused information centrality*, as defined in [3], which rank the nodes in the topic graph with respect to their *semantic centrality* to the seed concepts.

At this stage, each previously identified topic is represented by the set of seed concepts, the ranked label nodes, and the semantic network extracted at Step 4. This network contains some hundreds or thousands of nodes and edges. In the user interface, we present compressed topic graphs, consisting of only the shortest paths between seeds and labels as well as the nodes on these paths. Colour coding is used to differentiate the types of nodes and their relevance to the topic. Users can dynamically select how many of the top labels they want to inspect, resulting in graphs of different complexity.

3 Kanopy in Action

Kanopy is deployed as a web application. Users are encouraged to input any text in the user interface. Let us assume they copy and paste the body of a text related to research about the indian tigers endangered by extinction due to poor

genetic diversity [10]. In the following, we exemplify some key insights that can be gained with the current version of Kanopy [7].

Some topics that Kanopy identifies with the default settings are about locality (India), scientific research, and genetics. The third one is of particular interest. As the column `Extracted Concepts` shows, it brings together different concepts found in the text, such as "Preservation breeding", "DNA", "Gene pool", "Extinction" and "Genetic structure". Opening the topic graph and setting the `Top Labels` slider to 1 reveals that the top candidates for labelling are either *Genetics* or *Biology*, depending on the chosen `Centrality Measure`. Hovering over the topic graph highlights that *Genetics* is directly connected to the text mentions of "Genetic structure" and "DNA", a fact that explains its high score. The graph also clarifies that *Population Genetics*, the second-ranked label, is part of *Genetics* – which in turn is part of *Biology*. None of these recommended labels occured in the original text. Setting the `Topic granularity` to "Fine" in the user interface, Kanopy splits this topic into two more focused ones, labelled *Population Genetics* and *Conservation Biology*. Besides these multi-concept topics, Kanopy also displays single concepts that remained isolated, such as *Tiger*, together with the DBpedia categories and classes they belong to.

We plan to demonstrate Kanopy's research along the lines outlined above. A future extension we envisage for the system is that of corpus analysis and exploration. Here, Kanopy can be used to extract an interconnected network of concepts, topics and documents. This use case would bring value for a range of domains, such as online journalism, education, and knowledge management.

Acknowledgements. This work was supported by Science Foundation Ireland (SFI) partly under Grant No. 08/CE/I1380 (Lion-2) and partly under Grant No. 08/SRC/I1407 (Clique: Graph and Network Analysis Cluster).

References

1. L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: explaining relationships between entity pairs. *Proc. VLDB Endow.*, 5(3):241–252, Nov. 2011.
2. I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. An Eigenvalue-Based Measure for Word-Sense Disambiguation. In *FLAIRS '12*, 2012.
3. I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. Unsupervised Graph-based Topic Labelling using DBpedia. In *WSDM '13*. ACM, 2013.
4. G. Kasneci, S. Elbassuoni, and G. Weikum. Ming: mining informative entity relationship subgraphs. In *CIKM '09*, pages 1653–1656. ACM, 2009.
5. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *I-Semantics '11*, pages 1–8, 2011.
6. `http://dbpedia.org/`.
7. `http://kanopy.deri.ie`.
8. `http://nlp.stanford.edu/software/corenlp.shtml`.
9. `http://wikipedia-miner.cms.waikato.ac.nz/`.
10. `http://www.bbc.co.uk/news/uk-wales-22536571`.
11. `http://www.opencalais.com/`.

12. http://www.wikipedia.org/.
13. http://www.zemanta.com.