OCTVis: Ontology-based Comparison of Topic Models

by

Amon Dongfang Ge

B.Sc., University of British Columbia, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

***OCTVis*: Ontology-based Comparison of Topic Models**

submitted by **Amon Dongfang Ge** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Giuseppe Carenini, Computer Science
Supervisor

Dongwook Yoon, Computer Science
Supervisory Committee Member

Abstract

Topic modeling is a natural language processing (NLP) task that statistically identifies topics from a set of texts. Evaluating results from topic modeling is difficult in context and often requires domain experts. To facilitate evaluation of topic model results within communication between NLP researchers and domain experts, we present a visual comparison framework, *OCTVis*, to explore results from two topic models mapped against a domain ontology. The design of *OCTVis* is based on detailed and abstracted data and task models. We support high-level topic model comparison by mapping topics onto ontology concepts and incorporating topic alignment visualizations. For in-depth exploration of the dataset, display of per-document topic distributions and buddy plots allow comparison of topics, texts, and shared keywords at the document level. Case studies with medical domain experts using healthcare texts indicate that our framework enhances qualitative evaluation of topic models and provide a clearer understanding of how topic models can be improved.

Lay Summary

Topic modeling is a statistical approach that summarizes collections of text into topics being discussed. While topic modeling can be an effective way of understanding large text collections (often infeasible for a human to read through one by one), evaluating and improving these machine results requires human input and interaction. Particularly, domain experts can inform topic modeling results with the help of established domain knowledge, known as ontologies. We present an interactive visualization, *OCTVis*, to compare results from topic models within the context of a domain ontology. We evaluate the effectiveness of *OCTVis* with medical domain experts using online patient discussions and a medical ontology. Through ontology-based topic model comparison, *OCTVis* enhances qualitative evaluation of topic models and provides a clearer understanding of how topic models can be improved.

Preface

This thesis is the original work of the author, Amon Ge. All design, experiments, and writing were done by Amon Ge with help from Dr. Hyeju Jang and under the supervision of Dr. Giuseppe Carenini. Dr. Giuseppe Carenini and Dr. Dongwook Yoon provided feedback for revising the manuscript.

Table of Contents

Abstract . . . iii

Lay Summary . . . iv

Preface . . . v

Table of Contents . . . vi

List of Tables . . . viii

List of Figures . . . ix

Acknowledgements . . . x

1 Introduction . . . 1

2 Related Work . . . 3
2.1 Visualizing topic models for exploration . . . 3
2.2 Comparison of topic models . . . 6
2.3 Ontology visualization . . . 10

3 Data and Task Abstractions . . . 14
3.1 Data model . . . 14
3.2 Task model . . . 17
3.2.1 Evaluative metrics . . . 18
3.2.2 Tasks . . . 19

4 Solution . . . 24
4.1 Topic alignment . . . 24
4.1.1 Encodings . . . 24
4.1.2 Interactions . . . 26
4.2 Ontology mapping . . . 28
4.2.1 Encodings . . . 28

4.2.2 Interactions . . . 30
4.3 Document-centred comparison . . . 30
4.3.1 Encodings . . . 30
4.3.2 Interactions . . . 31
4.4 Key considerations during the design process . . . 31
4.5 Implementation . . . 36

5 Case Studies . . . 38
5.1 Methodology . . . 38
5.2 Results . . . 40

6 Discussion and Future Work . . . 43

7 Conclusion . . . 46

Bibliography . . . 47

A Pre-Study Questionnaire . . . 52

B Post-Study Questionnaire . . . 53

C Case Study Sample Dataset . . . 55

D Case Study Script . . . 56
D.1 Introduction and pre-study questionnaire . . . 56
D.2 User training . . . 57
D.3 Main study . . . 58
D.4 Post-study questionnaire and debrief . . . 58

List of Tables

5.1 Case study corpora statistics 38
5.2 Pre-study and post-study questionnaire responses 40

List of Figures

2.1 UTOPIAN from Choo et. al. 3
2.2 ParallelTopics from Dou et. al. 4
2.3 ThemeDelta from Gad et. al. 4
2.4 TopicPanorama from Wang et. al. 5
2.5 MultiConVis from Hoque and Carenini 5
2.6 SpecEx from El-Assady et. al. 6
2.7 Topic model comparison from Alexander and Gleicher 7
2.8 Comparison of papers from InfoVis, SciVis, and Siggraph from Oelke et. al. 8
2.9 Correspondence chart between latent topics and reference concepts from Chuang et. al. 9
2.10 Document relevance feedback view from El-Assady et. al. . . 10
2.11 (Part I) Schematic examples of ontology visualizations from Dudáš et. al. 11
2.12 (Part II) Schematic examples of ontology visualizations from Dudáš et. al. 12
2.13 PhenoLines from Glueck et. al. 13

3.1 A part of the UMLS semantic network 14

4.1 The *OCTVis* interface 25
4.2 The topic alignment facet 26
4.3 Topic alignment linked highlighting 27
4.4 The ontology facet 28
4.5 The document-topic heatmap 29
4.6 Topic coin clustering mockup 31
4.7 Whiteboard mockup of topic-centred and document-centred views 33
4.8 Mockup of topic alignment matrices, document-topic matrices, and document view 34
4.9 Colour binning for document keywords 35

Acknowledgements

I would like to extend my deepest gratitude to Dr. Giuseppe Carenini for his mentorship and kindness throughout my studies. This would not have been possible without his direction, wisdom, patience, and support.

Many thanks go to Dr. Hyeju Jang and Patrick Boutet for their eternal support since the start of my graduate studies. Their academic, social, and emotional guidance have gotten me to where I am today.

I am very grateful to Dr. Dongwook Yoon for his valuable feedback on the writing of this thesis.

Dr. Richard Lester's expertise and patience have been ever gracious, and I have learned a great deal from his advice. As well, thanks to Dr. Richard Lester, Dr. John-Jose Nunez, Dr. Kendall Ho, and Dr. Young Ji Lee for their time and intimate feedback during this project's case studies.

Most importantly, my love and my thanks to my partner, Alexandra Kim, and my family, for always believing in me.

Chapter 1

Introduction

Topic modeling is a natural language processing (NLP) technique that identifies topics in a set of texts, or corpus [6]. This technique is a critical component in most text analytics pipelines. The resulting topic models can provide informative and concise summaries of the content of large corpora, which can effectively support their exploration and analysis in diverse domains such as health care [7] and education [35]. However, evaluation of topic modeling results is challenging because no general gold standard is available, and automatic evaluation metrics, like topic coherence [39], at best weakly correlate with human judgments of quality [8, 10]. Thus, human-in-the-loop evaluation can be beneficial for improving the model, as shown in prior work [1, 10, 34].

Generally speaking, we need both domain experts and NLP researchers to make human-in-the-loop evaluation effective and efficient. Topic modeling for a particular text domain like health care requires domain expertise to evaluate whether topic modeling accurately captures the high-level overview of a corpus. Domain experts can interpret the results of a model given their domain knowledge, and NLP researchers can take that feedback to enhance the model. Arguably, the most critical step in this collaborative process is supporting the comparison of two different topic models. Whenever the NLP experts are faced with critical design choices in developing topic modeling solutions, they can show domain experts the resulting topic models and interactively gather their assessments of topic quality. Furthermore, topic models often involve large number of items (e.g., topics, phrases, words, documents) and relationships among them. The exploration and assessment of such complex models can greatly benefit from visual and interactive solutions [33].

In this thesis, we present a visual comparison framework for exploring results from topic models, which is designed to improve the communication between domain experts and NLP researchers in formative evaluation of topic models. Our framework support inter-model comparison by incorporating topic alignment visualizations. In addition, we introduce a novel ontology view, which provides a taxonomy of domain specific concepts (i.e.,

an ontology) for interpreting/labeling topics in a particular domain. The domain specific ontology can situate topic model comparison by mapping expected and understood high-level topics ("concepts") in specific text domains (e.g., medicine) to the actual results from topic models. For further in-depth exploration of the dataset, display of per-document topic distributions and buddy plots allow comparison of topics, texts, and shared keywords at the document level.

Our main contributions can be summarized three folds:

1. We introduce *OCTVis*, an analytic framework that allows for high-level and in-depth exploration and comparison of topic models;

2. Our framework provides a visual mapping of ontology concepts to topics on models generated from domain specific text, which is intended to enhance the interpretability of topic modeling results; and

3. Our use case studies in the health care domain demonstrate that our framework enhances qualitative evaluation of topic models and provides a clearer understanding of how topic models can be improved.

As a preview of the thesis, we discuss related work in visualizing topic models for exploration and for comparison, as well as visualizations of domain ontologies in Chapter 2. In Chapter 3, we derive data and task models for effective ontology-based topic model comparison. In Chapter 4, we present *OCTVis*, our visual framework, and describe the visual encodings in the final solution, considerations throughout the design process, as well as implementation details. In Chapter 5, we present the methodology and results from conducting case studies with domain experts on *OCTVis*. We discuss our evaluation as well as future work in topic model comparison in Chapter 6. Finally, we summarize our contributions in Chapter 7.

Chapter 2

Related Work

In the following sections, we address existing visualization research on topic modeling, topic model comparison, and situating topic models in their context via domain ontologies.

2.1 Visualizing topic models for exploration

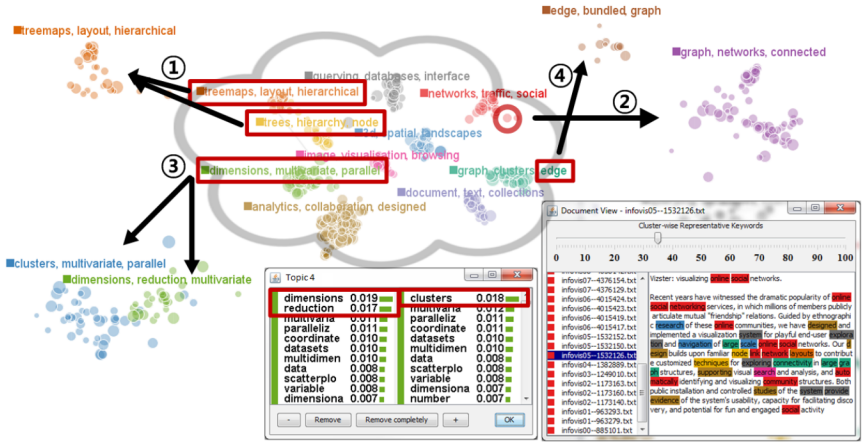


Figure 2.1: UTOPIAN from Choo et. al. [9]: (top) topic clusters; (bottom right) individual documents.

Text corpora and tasks for exploring their topic models are rich and diverse, as is the space of visual analytics designed for them. High-level *tag cloud* approaches [37] cluster topics and/or their keywords based on word similarity, displaying them in 2D space with size encoding keyword or topic strength (e.g. SolarMap [5], UTOPIAN [9] (Fig. 2.1), TopicPanorama [41] (Fig. 2.4)). Visualizations including UTOPIAN [9] (Fig. 2.1), ParallelTopics [12] (Fig. 2.2), iVisClustering [29], and ConVis [23] allow for in-depth exploration of document-topic distributions and documents, highlighting document keywords pertaining to particular topics. Visual river

metaphors seen in solutions including ThemeRiver [22], TIARA [30], HierarchicalTopics [13], and ThemeDelta [18] (Fig. 2.3), are commonly used to take into account the temporal nature of corpora (especially streaming data) and represent topics as they evolve over time.

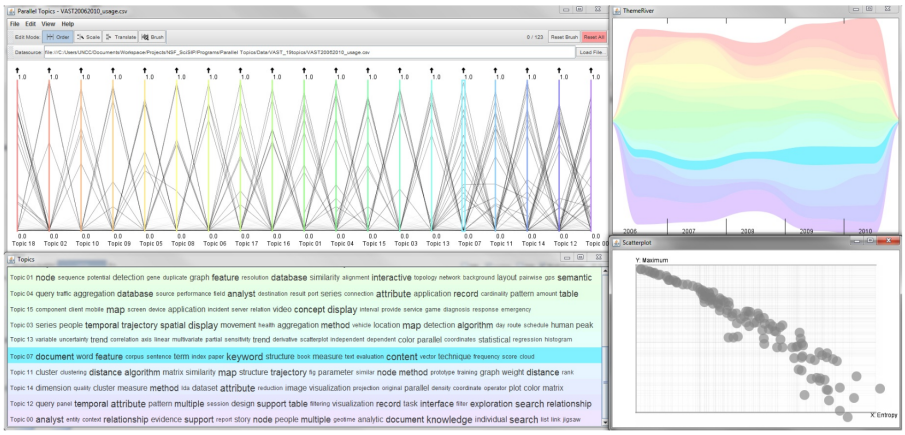


Figure 2.2: ParallelTopics from Dou et. al. [12]. Of note: (top left) parallel coordinate view of document-topic distributions; (top right) temporal river view of topics; (bottom left) word cloud of topic keywords.

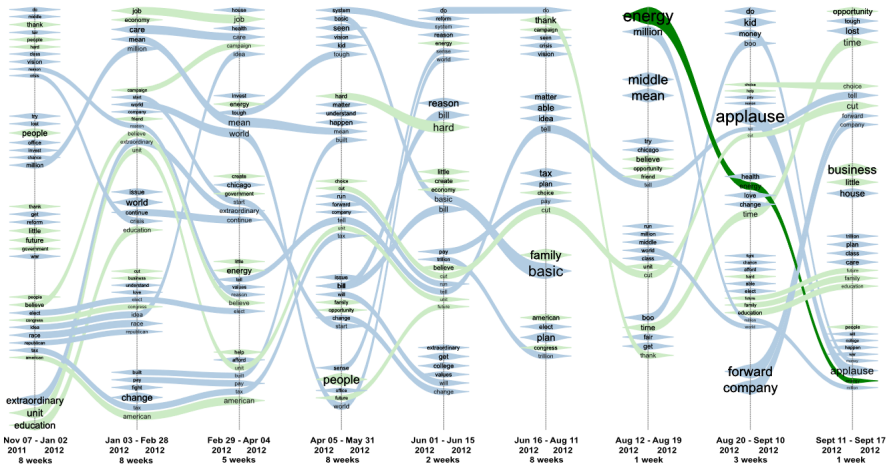


Figure 2.3: ThemeDelta from Gad et. al. [18] shows topics from Obama campaign speeches over the course of the US 2012 presidential election.

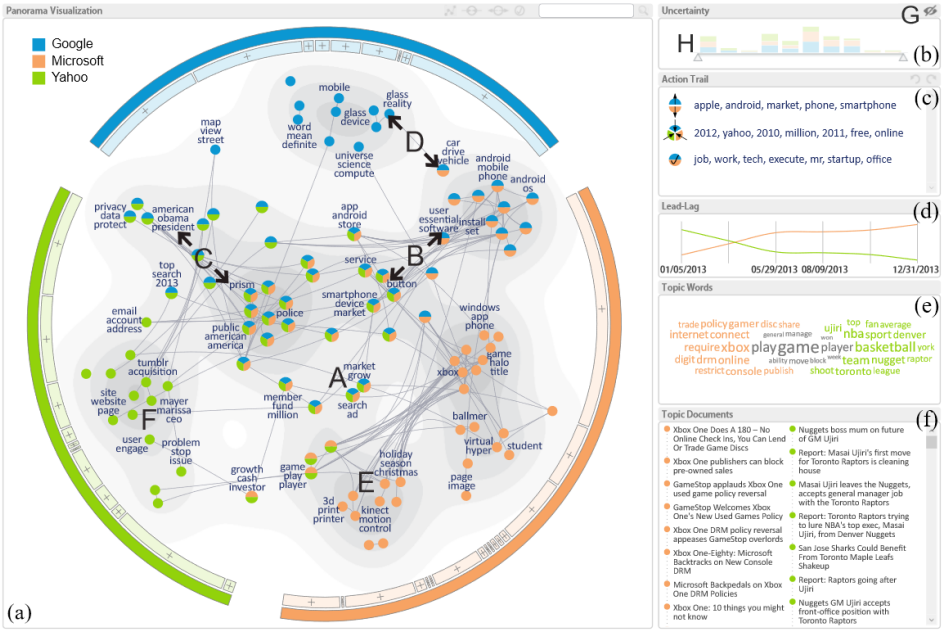


Figure 2.4: TopicPanorama from Wang et. al. [41] displays topics related to Google, Microsoft, and Yahoo. Of note: (a) hierarchical level-of-detail visualization of topics; (e) tag cloud of topic keywords; (f) documents related to the topic.

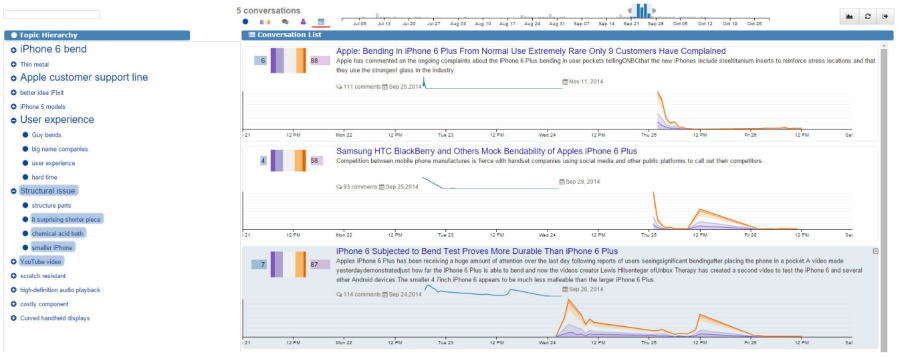


Figure 2.5: MultiConVis from Hoque and Carenini [24]. (left) Topic hierarchy; (right) Blog conversations with sentiment timelines.

In order to address text and topic models at scale, hierarchical topic modeling frameworks (e.g. HierarchicalTopics [13], TopicPanorama [41] (Fig. 2.4), MultiConVis [24] (Fig. 2.5)) group topics at varying levels of scale. More recently, interactive human-in-the-loop techniques including MultiConVisIT [25] and SpecEx [16] (Fig. 2.6) have shown that ongoing human intervention can dramatically improve the quality of topic models by adding, removing, merging, and splitting topics. Our work builds off of this existing topic modeling work using familiar visual idioms including tag clouds to display topic keywords and highlighting document keywords, in order to address comparison based tasks for exploring topic models with the specific focus of supporting domain and NLP experts in jointly comparing two topic models.

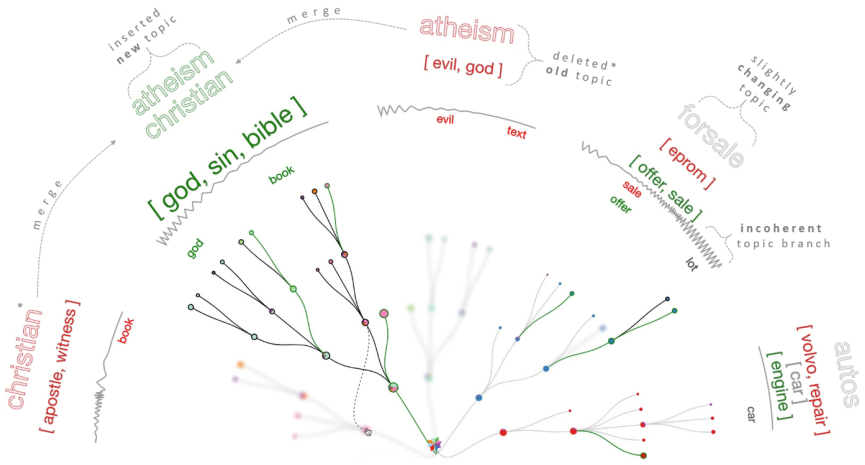


Figure 2.6: SpecEx from El-Assady et. al. [16] compares differences between two iterations of a topic model in the *Tree-Speculation View*.

2.2 Comparison of topic models

The task framework set by Alexander and Gleicher [1] breaks topic model comparison into three main tasks: topic alignment, distance comparison, and timeline comparison (see Figure 2.7). To address the task of topic alignment, the authors introduce bipartite graphs and heatmaps for high-level comparisons, as well as colour field rank comparisons for more in-depth topic-to-topic keyword comparison. To compare document distances, they introduce buddy plots: by holding one document stationary, relative dis-

tances to other documents in the corpus can be encoded with distance for one model and colour hue for the other. Sharp deviations in a smooth gradient indicate outlier documents with varied relative distance between the two models. For timeline comparison, asymmetric topic flow diagrams juxtapose two river flow visualizations [18, 22, 30] to facilitate comparison. While the authors introduce a number of useful techniques for high-level topic model comparison, our detailed task analysis indicates that intra-document exploration needs more support. Thus, our work extends their framework by adding in-depth document keyword comparison, ontology-based comparison. Furthermore, we add a critical step that was missing in their work. While Alexander and Gleicher [1] only showcased their interface on usage scenarios, we evaluate ours in several case studies with domain-specific corpora and domain experts.

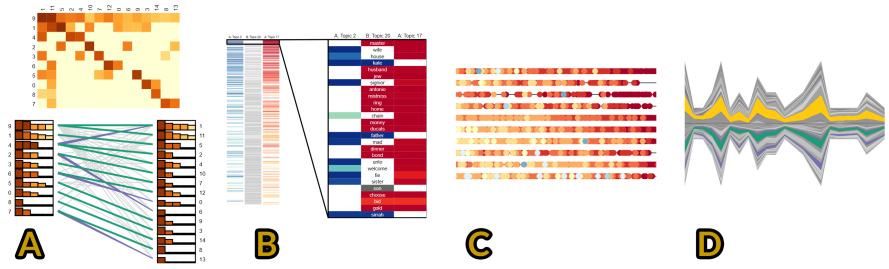


Figure 2.7: Topic model comparison from Alexander and Gleicher [1]. (A): Topic alignment heatmap and bipartite graph. (B): Colour field rank comparison between two topics. (C): Parallel buddy plots. (D): Asymmetric topic flow diagram.

Other comparative frameworks include Srinivasan et. al. [38], who thoroughly evaluated comparison tasks (not specifically for topic modeling) on bar charts and established a preference for redundant visual encodings (i.e. juxtaposition + explicit) for comparison. Our framework follows these principles to juxtapose results from two topic models and redundantly encode topic alignment with a parallel coordinates graph and a heatmap matrix.

Oelke et. al. [34] compare individual topics between supervised classes in one topic model using a 2D Euler diagram to represent class containment (see Figure 2.8). The authors use a *topic coin* glyph to display a word cloud of the top twelve keywords in each topic, as well as encoding topic distinctiveness and characteristicness based on other topics in the same class. While their multi-class comparison over topics is highly scalable, it requires well defined, supervised class labels over the corpus to compute topic discrimination. This

approach is very different from our framework, which compares results from two topic models and facilitates more interactive exploration of the corpus.

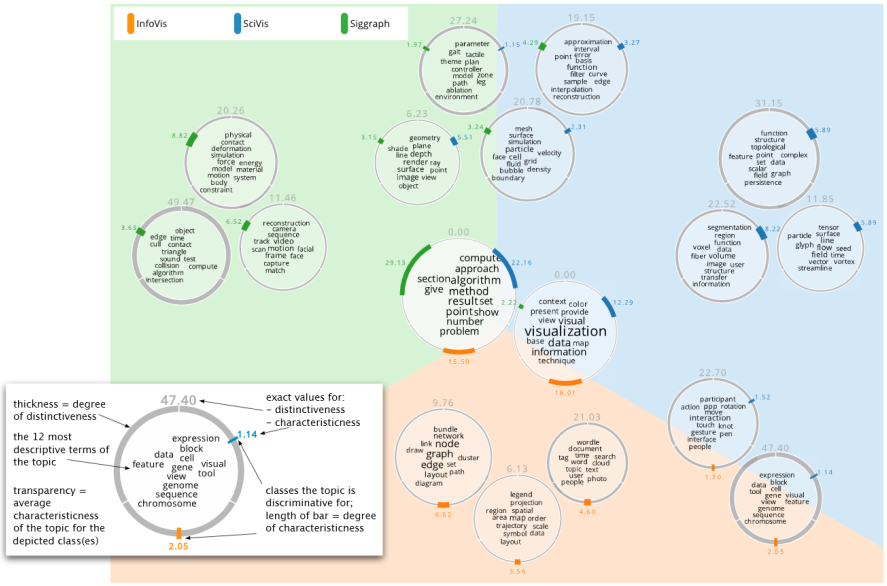


Figure 2.8: Comparison of 495 papers of InfoVis, SciVis, and Siggraph from Oelke et. al. [34]. Each circle is a topic, and its location encodes how discriminating that topic is for its class (the three conferences).

To compare topic model results with reference concepts, Chuang et. al. [10] present a probabilistic topic alignment framework. Reference concepts can come from domain experts, available metadata, or the outputs of other topic models. The authors compute alignment and four types of misalignments between the latent topics and reference concepts: *junk* topics (no matching concept), *fused* topics (2+ matching concepts), *missing* concepts (no matching topic), and *repeated* concepts (2+ matching topics) (see Figure 2.9). While this framework is a useful diagnostic tool for tuning LDA topic models via topic alignment, it expects a one-to-one correspondence between topics and concepts, which may not be the case when mapping topics onto a hierarchical ontology. Moreover, evaluation found discrepancies between experts and computed likelihoods when determining significant topics – many topics marked as significant were considered meaningless junk by experts. Our solution expands this approach by presenting an interactive visualization that allows users to *explore* the results of two models, which need not be probabilistic. Our mapping of domain-specific ontology onto topics does

not assume that the ontology is complete nor that the corpus covers all of the ontology, instead allowing open-ended comparison tasks in the context of the text domain.

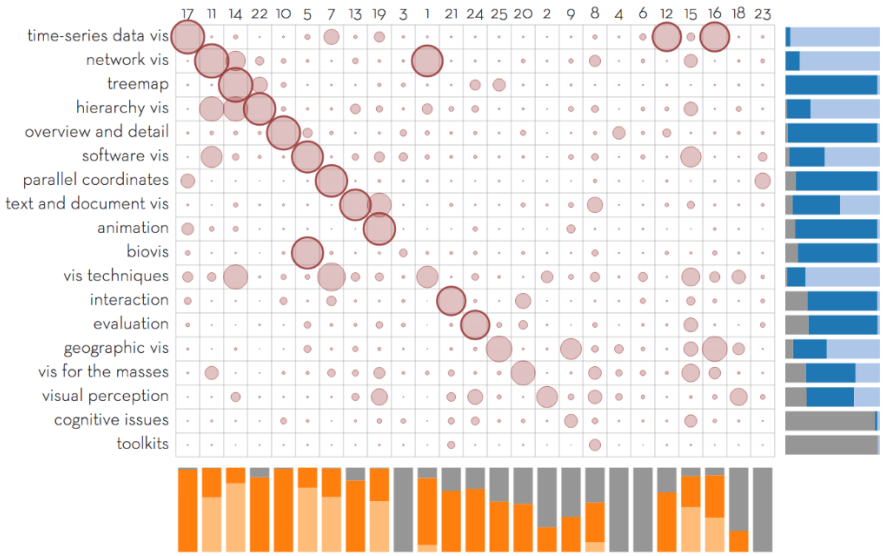


Figure 2.9: Correspondence chart between latent topics (columns) and reference concepts (rows) from Chuang et. al. [10]. Circle areas show matching likelihoods, while bars denote probabilities of concept/topic absence, one-to-one correspondence, and merging.

More recent interactive topic modeling frameworks [15, 16] provide an excellent foundation for comparing the results from a topic model against newer, re-parameterized results, while constantly responding to feedback from a human in the loop. El-Assady et. al. [15] facilitate a nuanced, in-depth exploration and comparison of the topic models and the corpus in two ways: comparative bar charts to juxtapose topic modeling parameters from two models side by side, and a *document relevance feedback view* to compare topic keywords between the models within certain documents (see 2.10). In order to preview the downstream effect of user changes to the model, SpecEx [16] incorporates several computed topic quality metrics including *separation*, *distinctiveness*, *coherence*, and *pointwise mutual information*. While these metrics may not correspond well with human judgments of quality [8, 10], they remain useful to estimate certain aspects of topic models. Our work incorporates computed metrics via juxtaposed bar charts as well

as supporting in-depth exploration of documents adapted specifically for comparison tasks between differing topic models.

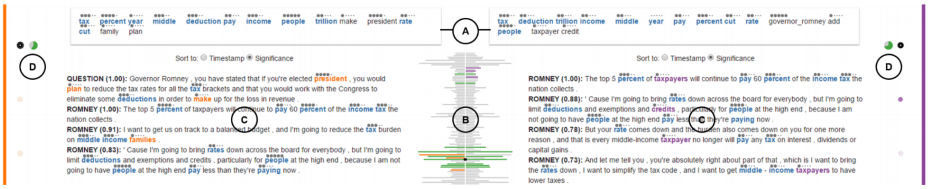


Figure 2.10: The document relevance feedback view from El-Assady et. al. [15] compares topic keywords over document snippets between two iterations of a topic (old: orange, new: purple, shared: blue). (A) Topic keywords of the two topics; (B) Documents in the corpus; (C) Top ten sentences for each topic; (D) Percentage of matching documents in the corpus.

2.3 Ontology visualization

Visualizations for *ontologies* – hierarchical knowledge taxonomies for particular text domains – have been explored extensively in previous work. Visual encodings have included the use of indented lists, space-filling techniques [14, 27], and node-link graphs [17, 27] (see Figures 2.11 and 2.12 for examples). To the best our knowledge, however, there is little prior work on visualizing ontologies for topic modeling comparison. The idea of *reference concepts* introduced by Chuang et. al. [10], and used to validate a single topic model is similar to what is captured by an ontology mapping. Unlike hierarchical ontologies, which often have high-level concepts that map to several topics, they expect an optimal one-to-one alignment. The PhenoLines tool [20] (Fig. 2.13) applies the topology of the domain-specific Human Phenotype Ontology to structure topic model output, but is built for comparing domain classes ("phenotypes") within subtypes of their topic model, instead of for comparing results from differing topic models. In this thesis, we use established list and graph visual encodings to map any domain-specific ontology against two topic model results, allowing general high-level comparison for any text domain with established taxonomies.

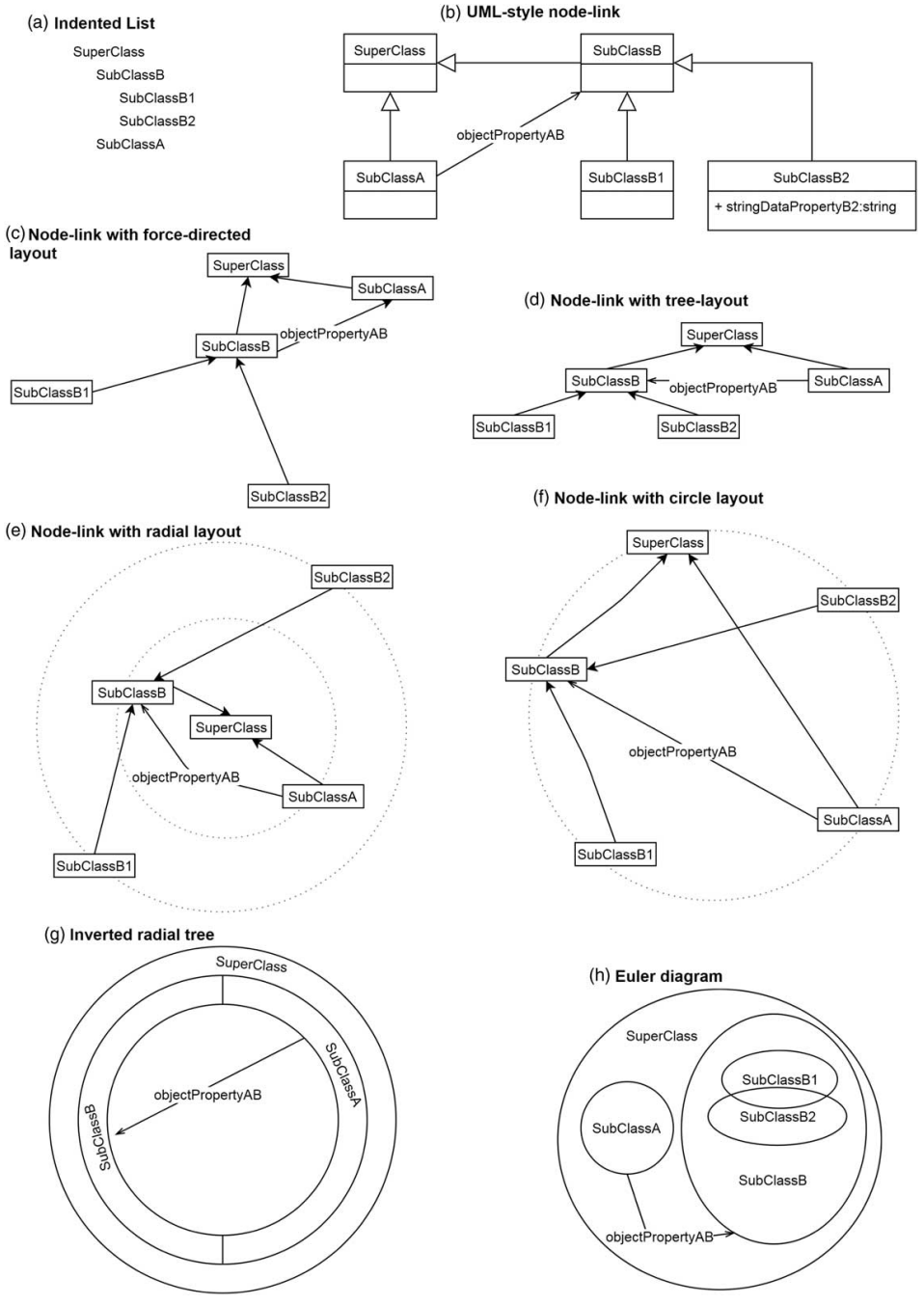


Figure 2.11: (Part I) Schematic examples of ontology visualizations from Dudáš et. al. [14]

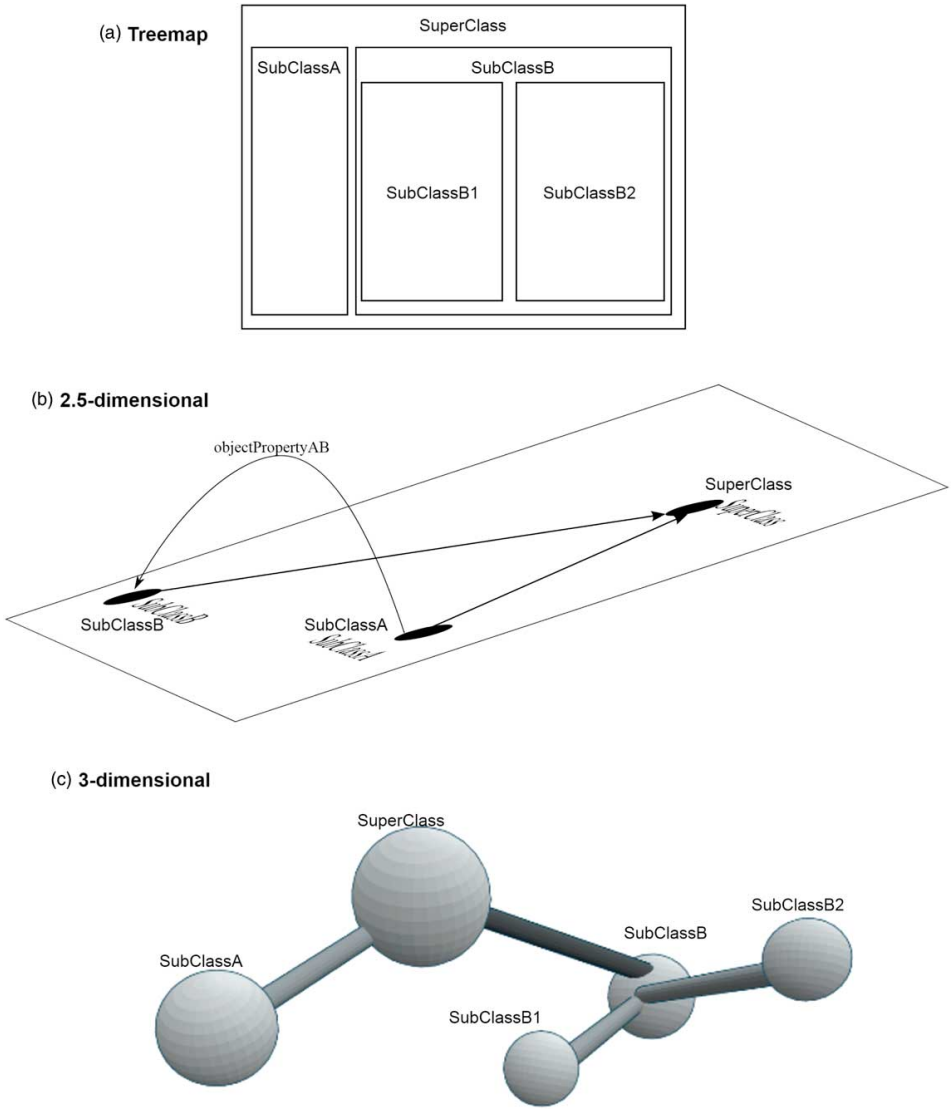


Figure 2.12: (Part II) Schematic examples of ontology visualizations from Dudáš et. al. [14]

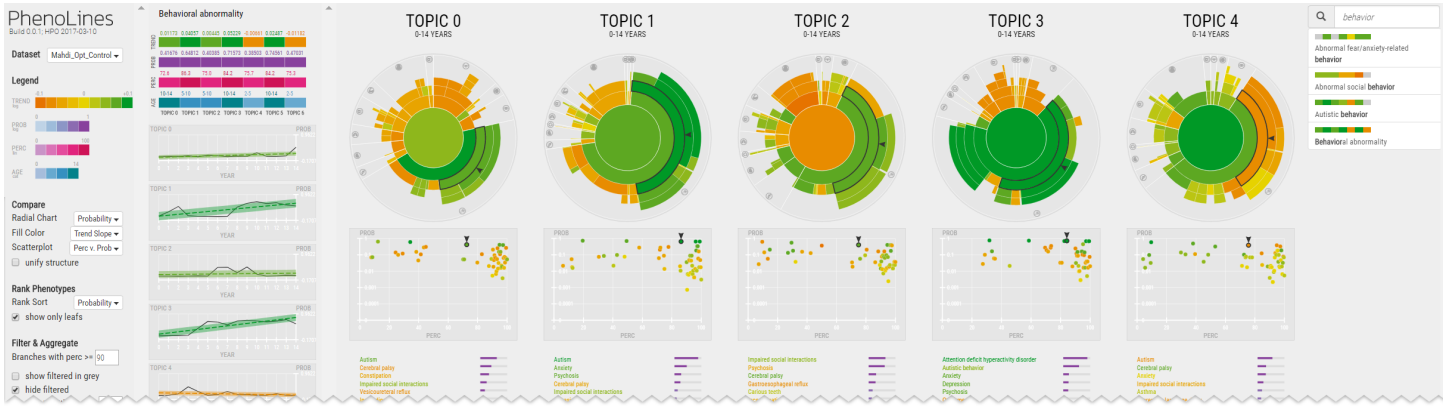


Figure 2.13: PhenoLines from Glueck et. al. [20], a visualization for optimizing topic models that describe disease symptoms using the Human Phenotype Ontology.

Chapter 3

Data and Task Abstractions

By following the standard information visualization methodology [32], we base the design of our visual interface, *OCTVis*, on abstracted data and task models. The data model describes text corpora, topic models, and ontologies, while the task model outlines key comparison tasks to enhance qualitative evaluation and understanding of ontology-based topic models.

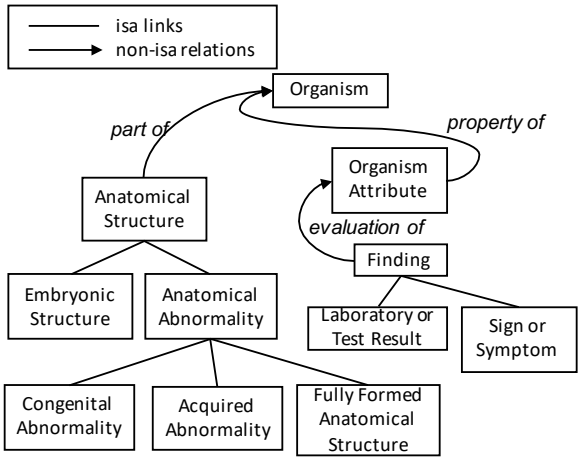


Figure 3.1: A part of the UMLS semantic network.

3.1 Data model

OCTVis is designed to support the comparison of two different topic models for the same corpus. At the highest level, the data model includes: (i) the set of documents comprising the corpus, (ii) two topic models, and (iii) a domain specific ontology.

Text often comprises of many topics. Topic modeling is a key NLP task that automatically identifies topics within a particular text document or corpus. For instance, a news article about electric vehicles could cover battery

technologies, history of electric vehicles, the current market and competitors, government incentives, the environment, and so on. In conversational text, like email threads and online discussions, topics can vary drastically within a single thread or discussion, and can also include subjective content such as emotion and sentiment. Identifying topics within a text is crucial for downstream visual analytic tasks like text comparison and exploration.

A topic model dataset can be represented as multiple tables. One table stores topics (attributes) per document (row), assigning a quantitative value representing the strength of that topic for that particular document [33]. For probabilistic methods like Latent Dirichlet Allocation (LDA) [3], each value in this table is the specific probability, $P(t|d)$, that the attribute topic t (column) occurs given the document d (row). For each document, we can also assign topics to subdivisions of text, most commonly per word. Over the entire corpus, we have a table where each row represents a topic, and each column represents a word in the corpus, and the value in each cell is a quantitative weight for that word in that topic (in LDA, this is the probability of a word w occurring given a particular topic t, $P(w|t)$).

Although most topic modeling techniques automatically generate labels for the extracted topics (e.g., most likely words from the topic in LDA), there often exist human-annotated labels for topics and subtopics within particular lexical domains, known as *ontologies*. These ontologies are hierarchical networks of categories and relationships that classify terms used in a particular domain. As an example, the Unified Medical Language System (UMLS) Semantic Network [31] is an ontology of terms, hierarchical categories ("concepts"), and relations used in the medical lexicon. Concepts within UMLS include "Laboratory or Test Result", "Sign or Symptom", and "Finding", and they are linked with their parent concepts by semantic relations, like `part_of` and `is_a` (i.e., subclass), as shown in Figure 3.1. In our work, we use the UMLS Semantic Network as our domain-specific medical ontology.

Algorithm 1[1]shows the pseudo-code to derive the data model, assuming that the two topic models to be compared have been already generated. This process involves the four steps below.

First, each model is mapped into the ontology. The step comprises two phases: (i) words in the corpus are mapped into the ontology, (ii) each topic of the topic model is mapped into ontology concepts. The first phase is

[1]All the procedures are here specified for an LDA model with conditional probabilities $P(t|d)$ and $P(w|t)$. For non-LDA topic models, corresponding weights connecting topics to documents and words to topics should be used.

Algorithm 1 Pseudocode for derived data: ontology mapping, topic alignment, an example of a topic metric computation (topic segregation), and document distances

procedure MapOntologyConceptsToTopics
    for each word w_j in the corpus **do**
        assign w_j to a concept if a mapping exists
        (based on an ontology labeler, MetaMap for the UMLS in this thesis)
    for each concept c_i **do**
        for each topic t_k (in both models) **do**
            $weight(t_k, c_i) \leftarrow 0$
            for word w_j mapped to c_i **do** ▷ see Equation 3.1
                $weight(t_k, c_i) \leftarrow weight(t_k, c_i) + P(w_j|t_k)$
procedure ComputeTopicAlignment
    for each topic t_j in model A **do**
        for each topic t_k in model B **do**
            $alignment = sim(t_j, t_k)$ ▷ see Equation 3.2
procedure ComputeTopicSegregation
    for each topic t_j in model X **do**
        for each topic t_k in model X **do**
            compute $sim(t_j, t_k)$
        $rawSegregation \leftarrow$ mean $sim(t_j, t_k)$ for all other t_k
    $segregation \leftarrow$ normalized $rawSegregation$ to sum to 1 over model X
procedure ComputeDocumentDistances
    for each document d_i in the corpus **do**
        for each document d_j in the corpus **do**
            $distance(d_i, d_j) \leftarrow sim(p(t|d_i), p(t|d_j))$

done by using MetaMap [2], a toolkit that automatically codes text with the UMLS semantic concepts. For instance, the text "I need running clothes" would be coded as "I need running `[Daily or Recreational Activity]` clothes`[Manufactured Object]`". In the second phase, to compute the strength of a mapping between a topic t_k and an ontology concept c_i, we use the equation:

$$weight(t_k, c_i) = \sum_{w_j \text{ mapped to } c_i} P(w_j|t_k), \tag{3.1}$$

where w_j is a word in the corpus, $P(w_j|t_k)$ is the probability of that word given the topic t_k.

Second, following Alexander and Gleicher [1], we compute a mapping (i.e., alignment) between the two topic models as the cosine similarity between their probability distribution of a word occurring given a particular topic, $p(w|t)$.

$$sim(t_j, t_k) = \frac{\sum_i P(w_i|t_j)P(w_i|t_k)}{\sqrt{\sum_i P(w_i|t_j)^2}\sqrt{\sum_i P(w_i|t_k)^2}}, \tag{3.2}$$

where $w_i \in V$ is each word in the corpus.

Third, we derive quantitative topic metrics per topic, including topic segregation (how well separated topics are), topic cohesion, and topic stability. These will be described and justified in the next section.

Finally, we compute document distances within each topic model. For each pair of documents in the corpus, we compute their relative distance using the cosine similarity between their document-topic distributions (see the ComputeDocumentDistances procedure in Algorithm 1).

3.2 Task model

The high-level goal of our visual solution for comparing topic model results is twofold:

- first, to assist communication between an NLP researcher and a domain expert during formative evaluation of topic modeling methods, and
- second, to assist an NLP researcher in the development of these novel techniques.

The process of refining this high-level goal into a hierarchy of user analytic tasks to drive the interface design was based on principles from existing literature on visualizing comparisons [1, 19] and topic modeling analytics [4, 11, 16], as well as an informal collection of user requirements from NLP and domain experts (including the authors).

In essence, the interface needs to support the assessment of *which of two given topic models is a better model for the target corpus*. This assessment can be decomposed into four key qualitative and quantitative metrics for evaluating how good a topic model is. We first describe these metrics and informally explain how users can be supported in evaluating those. Next, we provide a more formal task model for the metrics.

3.2.1 Evaluative metrics

Topic segregation This evaluative metric relies on the intuition that an ideal topic model should have different topics with little word or semantic overlap. While word overlap is easy to compute numerically, semantic overlap is a more difficult problem. The interface should allow the users to gauge the semantic distance between topics qualitatively, by looking for duplicate words, synonyms and similar/related words shared across topics.

Topic coherence The quality of a topic model depends on the quality of each of its topics. Thus, we also consider the intuitive metric of within-topic coherence. Essentially, words within a topic should share some overall semantic relatedness. A topic with seemingly "random" or "out of the blue" words is considered noisy, and having poorer quality than one with more coherent words. This idea relates to well defined quantitative metrics, including topic coherence[11], normalized pointwise mutual information[4], and topic stability[16].

Topic model comprehensiveness An important quality metric for a topic model is how thoroughly it covers the topics appearing in the corpus. For instance, suppose a corpus discusses the topics of 'food therapy', 'symptom', and 'sleep'; topic model A captures only 'food therapy', 'symptom', and other irrelevant words while topic model B captures all three topics successfully. In this case, model B would be preferred to model A since it is more comprehensive. This is very difficult to measure without a human-centred gold standard because it requires understanding the entire corpus. Our visual analytics solution aims to help researchers create a mental topic model to compare topic model results against, enabling evaluation

of model comprehensiveness. One approximation for this measure can be derived from mapping the topic model on to the ontology space, and seeing the topic coverage on the ontology concepts. While the ontology might not accurately reflect the corpus space, it may help the NLP researcher and the clinician create this mental model to evaluate model comprehensiveness and compare the comprehensiveness of different models.

Topic assignment quality A topic model should also map well both to ontology concepts and at the document level, in the sense of creating a meaningful middle level between the two. An ideal topic model is well situated both within the context of its text domain (and should map well to an ontology) and of individual documents (i.e. topics should reflect the actual text in each document). If a topic model has well separated high-level topics that are closely related to the domain ontology, but these topics do not clearly convey the actual points of discussion in the documents, the model is too high-level and not much more useful than the ontology itself. On the other hand, if topics are too fine-grained and describe the documents in detail without merging similar topics or summarizing key ideas, the model is not a useful substitute for the documents.

3.2.2 Tasks

With these evaluative metrics in mind, we provide a hierarchy of user analytic tasks for comparing ontology-based topic models.

Evaluating Overall Topic Model Quality (TM): These tasks cover the topic models overall, allowing for high-level comparison between the two models before and after delving into individual topics, ontology mappings, and documents. These comparisons should be considered throughout the exploration of the topic models.

- (TM1) *Overall, does one model describe the corpus better than the other model?*

  An overall quality comparison may be difficult to conclude, but this principal task prompts further investigation and guides the exploration of the topic models.

- (TM2) *Is one model more coarse-grained or fine-grained than another?*

  As mentioned earlier, an ideal topic model should be situated between the high-level ontology and the individual documents of the corpus.

Whether or not a topic model is at the right level of detail is relative (to the other model, to the ontology, and to the corpus), use-case dependent, and drives further comparison tasks.

- (TM3) Topic coverage/comprehensiveness/segregation:
  - (TM3a) *Are there topics that might be better divided into subtopics, and are there topics that might be better merged into supertopics?*
    Similar to TM2, certain topics may be too broad or specific, and topic overlap may suggest potential to merge topics. Proper or improper topic granularity is an indicator of overall model quality.
  - (TM3b) *How well are topics separated from one another?*
    This task overlaps with the previous task (TM3a) in comparing relative topic model quality. Distinct, well-segregated topics suggest a clearer overall model.

Assessing Topic Alignment (TA) A principal concern of topic model comparison is alignment between topics across models [1]. The following tasks address this need.

- (TA1) *How well do topics match? In particular, which topics are the most similar across models? Which topics are the most different?*
  For instance, in Figure 4.2, we see that topics A2 (keywords "eat like carbs don fat sugar think") and B1 (keywords "eat like carbs know don try good") have the highest pairwise alignment, while A4 ("study blood learned"...) and B2 ("juice green recipe"...) are the most dissimilar.
- (TA2) *Of a highly aligned topic pair,*
  - (TA2a) *How are the topics similar, and how are they different?*
    From the above example, we see that both topics cover similar keywords relating to diet and carbs.
  - (TA2b) *Based on this, which of the two topics better captures a topic actually covered in the underlying corpus?*
    While topic keywords may inform the quality of the topic, the users may also want to explore documents with these aligned topics and compare document keywords between them (see tasks DW1-3).

Verifying Topic Quality (TQ) These tasks involve evaluation of the quality of individual topics relative to one another within a single topic model and across topic models. As such, computed quantities including topic coherence and segregation may provide estimates for human qualitative measures of topic quality as well as provide points of comparison for understanding how and why topic models differ.

- (TQ1) *Do measured metrics agree with human judgment for topic quality?*
  - For example, a topic with high segregation (little semantic overlap) may be a good quality topic in the context of the corpus if it represents a clear and unique topic. On the other hand, a highly segregated topic may consist of nonsensical words that appear in few documents and be relatively uninformative.
- (TQ2) Coherence: *How coherent or noisy is a topic?*

  We can use both computed coherence and the topic keywords in a topic to evaluate the relative noisiness of a topic, a marker for topic quality.
- (TQ3) Topic coverage/comprehensiveness: *How well do individual topics cover the corpus?*

  Whether a topic spans the entire corpus or only shows up in a few documents can prompt further investigation.

Assessing Ontology Mapping Quality (O): The mapping from topics to a domain ontology situates each model within the context of the text domain it belongs to. Here, we outline informative tasks that compare topic models taking into account this ontology context.

- (O1) *Which ontology concepts are the most frequent in a topic?*

  For instance, if a topic is strongly mapped to the UMLS concepts "Pharmacologic Substance" and "Clinical Drug", we expect it to cover keywords and documents relating to drugs and medicine.
- (O2) *Which concepts are prevalent across both topic models?*

  Concepts covered by both models should suggest alignment of topics (see tasks TA1-2) and lead to similar comparisons of topic quality across models.

- (O3) *Which concepts are seen differently across topic models, and across topics, and why?*

  If an ontology concept is strongly mapped to one model's topics but not the other model's topics, it can be a red flag to investigate further differences among these topics. This may indicate more coherent topics, more segregated topics, or a missed topic entirely in one model.

- (O4) *Are there concepts which are more informative than others in describing the corpus?*

  This task informs the quality of the ontology itself which can help develop better ontology mapping techniques. For example, if users identify concepts such as "Biologic Function" that might be more informative if broken down into "Physiologic" and "Pathologic Function", updating our ontology mapping in our framework to reflect this feedback will better facilitate ontology-based comparisons.

- (O5) *Do these concepts make sense in the context of the topic words and overall corpus?*

  Ontology mappings need to be related back to the context of the corpus itself. High-level topics that map well to ontology concepts may not correspond to clear relationships at the document level, and prompts further investigation.

Comparing Document Distances (DD) Tasks for comparing document distances are key to understanding similarities within and across topic models [1]. As we saw in Chapter 3.1, given a topic model a distance measure between any two documents can be computed. The following tasks address these comparisons in document distances and can be applied globally (how documents cluster differently between overall topic models) and locally (how documents differ relative to a specific document of interest).

- (DD1) *How do the topic models cluster documents differently with respect to the distances among documents?*

- (DD2) *Are there documents that are similar across the models, and documents that are quite different?*

Exploring the Document Topic Distribution (DT) In-depth exploration of documents requires first understanding how topics are distributed across documents and comparing these distributions across models. These

tasks can identify key topics and documents of interest in the context of the text corpus, facilitating fine-grained comparison. A topic that spans most documents in the corpus may be less or more cohesive or informative than an infrequent topic, and topic models may perform differently for these kinds of topics. One model may have broad, spanning, and overlapping topics that cover the corpus thoroughly, while another may have sparse and well-segregated topics that pick out specific documents while having gaps in coverage.

- (DT1) *How are topics distributed among documents?*
- (DT2) *Are there topics that are prevalent throughout the corpus?*
- (DT3) *Are there topics that only show up in a few documents?*
- (DT4) *Are topics that are prevalent across the corpus substantially more or less informative than infrequent topics?*

Comparing Document Keywords (DW) At the finest level of comparison, we need tasks involving topic keywords at the document level to form an in-depth understanding of each topic. These tasks compare aligned topics across models by exploring the keywords captured by each topic within particular documents. High quality topics should maintain coherence at this level, and the keywords within a document should align well with the overall keywords of the topic.

- (DW1) *How well do the keywords in aligned topics match at the document level?*
- (DW2) *Does one model find keywords that another model misses?*
- (DW3) *Are topics assigned correctly at the document level?*

Chapter 4

Solution

Our proposed solution, *OCTVis* (Ontology-based Comparison of Topics), uses two linked views to assist in all the topic comparison tasks captured by our task model (see Figure 4.1). To support our topic-centred tasks, a high-level topic-centred facet presents the topics from two topic models to compare topic alignment, as well as the mappings from an existing ontology's concepts to each topic. For more in-depth, document-centred tasks, a document-centred facet shows topic distributions per document, relative document similarities, as well as supporting exploration of each individual document.

4.1 Topic alignment

4.1.1 Encodings

The topic-centred visualization facet, shown in Figure 4.2, builds on the similarity heatmap and parallel-coordinates node-link solution presented by Alexander and Gleicher [1]. Two topic models are *juxtaposed* side-by-side, using a standard word cloud encoding to show topic keywords, with their strength encoded by font size [37]. A parallel-coordinates facet in between the two models shows alignment between topics with colour, line width, and opacity all redundantly encoding the alignment strength. Below, a heatmap also shows this topic alignment information in a similarity matrix, with the topics of one model along one dimension and the topics of the other model along the other dimension. Fields are coloured by the strength of inter-model alignment between those two topics. Additional computed per-topic metrics, such as topic coherence or segregation, are encoded with aligned horizontal bars beside each topic node.

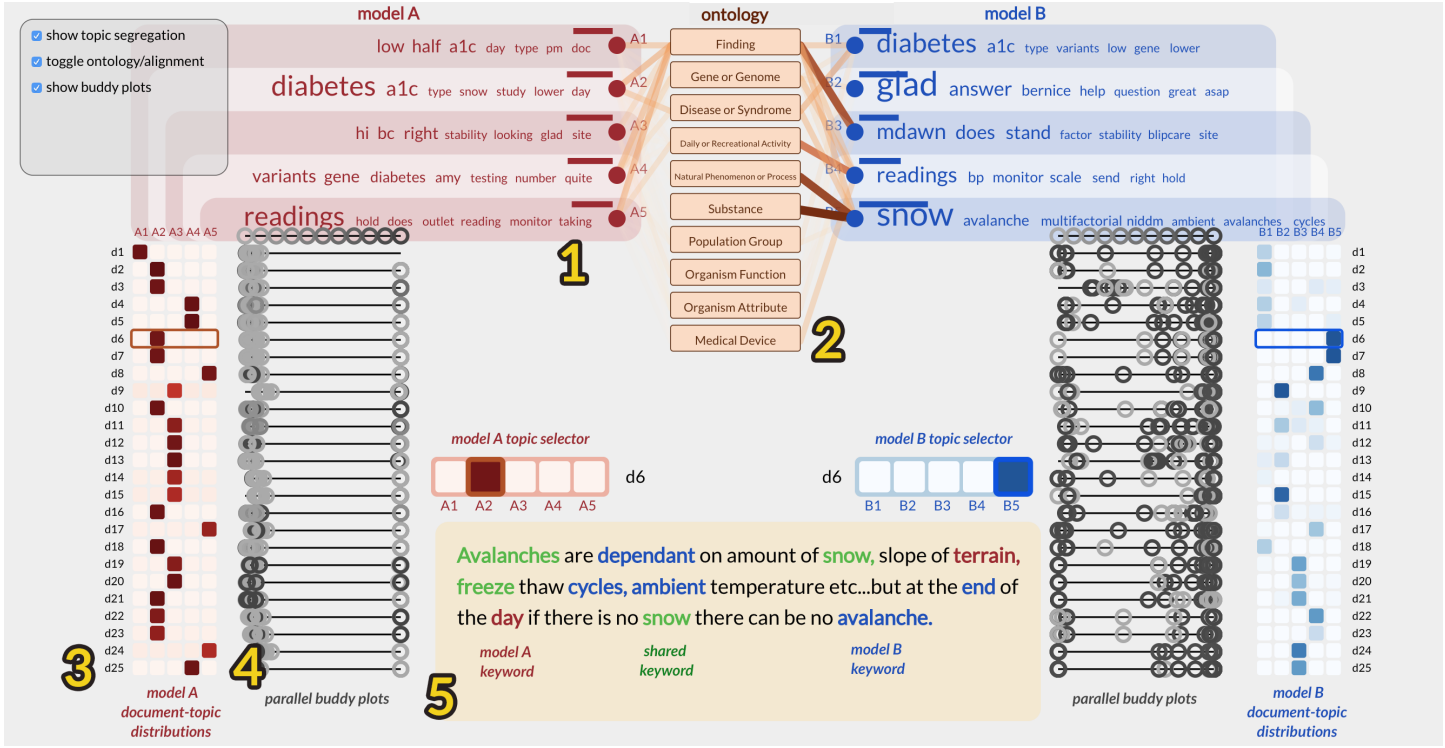


Figure 4.1: The *OCTVis* interface. Topic models (1) are juxtaposed side by side, allowing comparison of topic alignment to a **domain ontology** (2) and of document-topic distributions (3). Parallel buddy plots (4) compare relative distances between documents across both models. A document's keywords (5) of a selected pair of topics are highlighted if they belong to the **left model's topic**, **right model's topic**, or are **shared**.

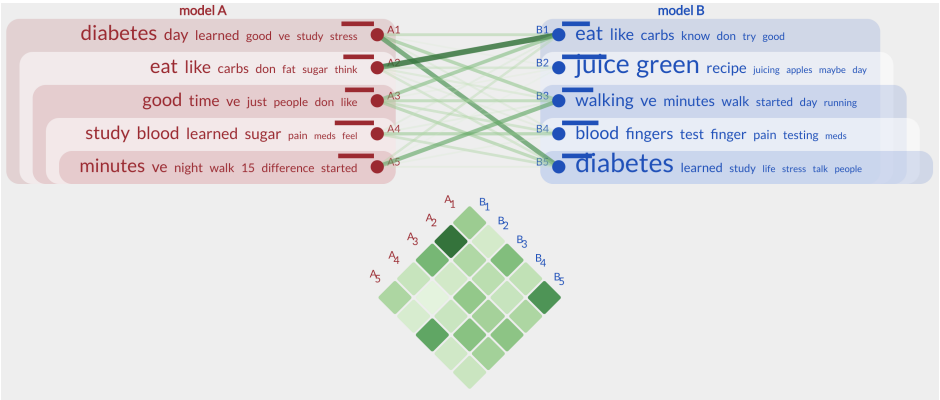


Figure 4.2: The topic alignment facet: topic keywords of two topics models are juxtaposed side-by-side, with model A on the left and model B on the right. The topic alignment node-link graph and heatmap (center) encodes topic alignment strength between pairs of topics across topic models.

Showing topic keywords beside topics supports the tasks of evaluating the quality of matched topics: users can determine how well two topics match (TA1) based on the computed alignment scores, compare the quality of the match using the topic keywords (TA2a, TA2b), as well as compare similarities and differences across model topics, between computed matches and non-matches (TA1). As well, users can evaluate individual topic quality (TQ1-3) based on their understanding of the topic given the keywords, computed per-topic metrics, and through comparison between matched topics. For instance, users can use the topic keywords along with the computed metric bars to evaluate whether these measured metrics agree with human qualititative measures (TQ1).

The topic alignment heatmap presents a higher-level overview of the alignment, supporting tasks including how well topics are segregated (TQ3) and which topics are most similar and most different (TA1-2). In combination with the topic alignment graph and topic keywords, this supports topic quality tasks including whether topics might be better subdivided or merged (TQ2), and in general whether one model is more coarse-grained or fine-grained than another (TM2).

4.1.2 Interactions

Users can filter the displayed computed metrics beside each topic node to compare metrics of their interest. For instance, while evaluating the co-

herence or noisiness of a topic, users can hide all metrics except for topic coherence.

As well, highlighting a heatmap field cross-filters that link in the parallel-coordinates facet, desaturating the other links and linking this highlighted link and corresponding topic nodes (see Figure 4.3). This supports alignment tasks that compare topics with high computed alignment (TA2a, TA2b), as well as quality evaluation at the topic level (TQ1-3) and at the model level (TM1-2).

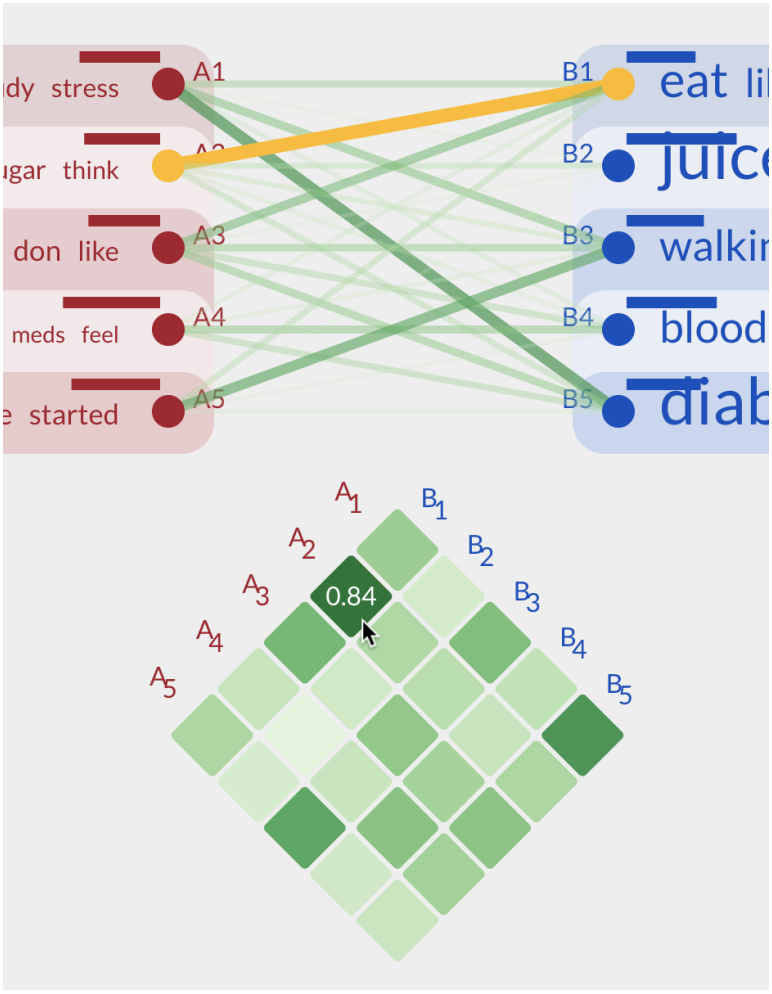


Figure 4.3: Linked highlighting between the topic alignment node-link graph and heatmap.

4.2 Ontology mapping

4.2.1 Encodings

The ontology mapping replaces the topic alignment parallel-coordinates graph and alignment heatmap components with a vertical list of ontology concepts (see Figure 4.4). A node-link graph on each side of the list connects each ontology concept with each topic within that side's topic model. Similar to the topic alignment graph, the strength of each alignment between ontology concept and topic (see Chapter 3.1) are redundantly encoded with colour, line width, and opacity. Concepts are sorted in descending order of combined strength from both topic models, so more frequent concepts are higher on the list than less frequent concepts. Topic keywords and per-topic metrics are still presented beside each topic node, as in the alignment view.

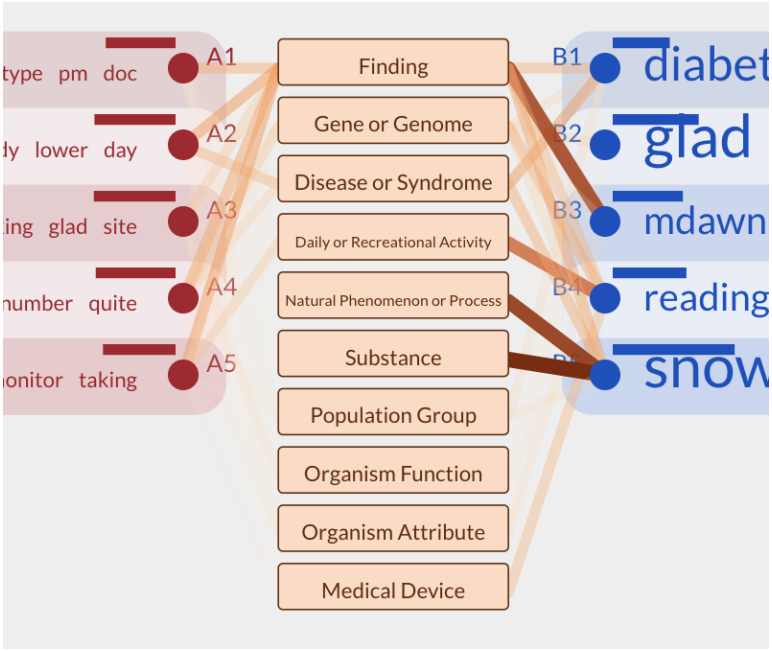


Figure 4.4: The ontology facet (orange, center) displays ontology concepts and their mapping to each topic. The weight of the mapping is encoded by the saturation, opacity, and width of the line.

The order of the concept list and the concept mapping supports ontology-related comparisons like finding the most frequent concept per-topic (O1) and finding prevalent concepts within both or one topic model (O2, O3).

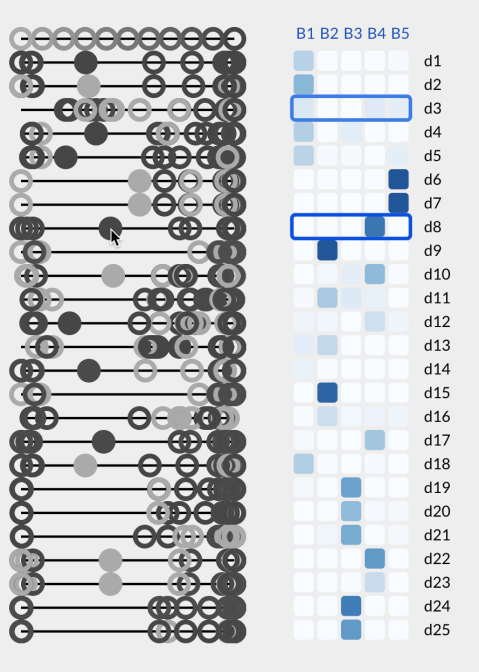


Figure 4.5: The document-topic heatmap (right) shows the distribution of topics (columns) for each document (rows) in this topic model. Strengths of topics are encoded by the saturation and lightness of the topic model's colour (red for the left topic, blue for the right topic). The buddy plot facet [1] (left) compares relative document distances across both models. For each document, relative distances of every other document (circles) within the adjacent model (blue) are encoded with horizontal position, with closer documents to the right. Relative distances for the opposite model (red, not shown) are encoded with lightness, with closer documents lighter than farther documents. For instance, most documents, except for one, are rather close to `d25`, according to the blue model, but this is not the case with respect to the other model (most circle are dark). Hovering over a document circle highlights that document in all the parallel buddy plots, and that particular row in the document-topic heatmap.

Using the topic mappings and topic keywords, users can then evaluate the informativeness of these concepts (O4) as well as whether or not they make sense within the context of the corpus (O5).

4.2.2 Interactions

Users can toggle between the topic alignment facet and the ontology mapping facet to support different tasks. Currently, *OCTVis* does not support ontology specific interactions, like showing more or less concepts. Such interactions, when appropriate, could be added in future versions (see Chapter 6).

4.3 Document-centred comparison

4.3.1 Encodings

Below each topic model's topics, we display a document-topic heatmap representing the distribution of topics within each document in the corpus, expanding on prior work in document-centred comparison [1] (see Figure 4.5). Each row represents a single document, and each column represents a topic from the above model. Within the heatmap, the saturation and lightness of the field encodes the strength of that topic (column) within that document (row). A checkerboard background hint relates the mapping from each column to each topic's keywords (See Figure 4.1 for these background hints).

To support the comparison of document distances (DD1, DD2), we adapt the parallel buddy plot encoding introduced by Alexander and Gleicher [1]. On each side of a topic model's document-topic heatmap, a buddy plot facet can be toggled to compare relative document distances in both models (see Figure 4.5). Each line corresponds to a document, and the circles represent every other document's relative distance to the reference document. Inter-document distances in the adjacent model are encoded by horizontal distance, while distances in the opposite model are encoded by lightness; if the two models have similar document distances, each buddy plot line would follow a smooth gradient like the reference line at the top of Figure 4.5. Sharp changes in this gradient indicate documents that have differing distances across the two models [1]. In *OCTVis*, we chose to use hollow circles instead of solid disks to see overlap between documents more clearly, and replaced hue encodings with lightness to prevent overloading established colour encodings for the two topic models.

4.3.2 Interactions

To explore the topic models at the document level, single documents are displayed in the centre of the interface, similar to interactive topic modeling techniques presented by El-Assady et al. [15]. Selecting a row in the document-topic heatmap displays that relevant document, in addition to a zoomed view of that document's topic distribution heatmap across each model (just above the document panel, see Figure 4.1). Clicking a particular field in the zoomed heatmap selects that topic for intra-document exploration. The user can select one topic from each of the two models and the strongest keywords for the selected topic are highlighted in the document view. Keywords belonging to the left topic are displayed in red, while keywords from the right topic are displayed in blue. Keywords shared across both topics are displayed in green.

4.4 Key considerations during the design process

In this section, we discuss key issues that we considured during the design process, including alternative views, layouts, and encodings that were not eventually chosen for the final solution.

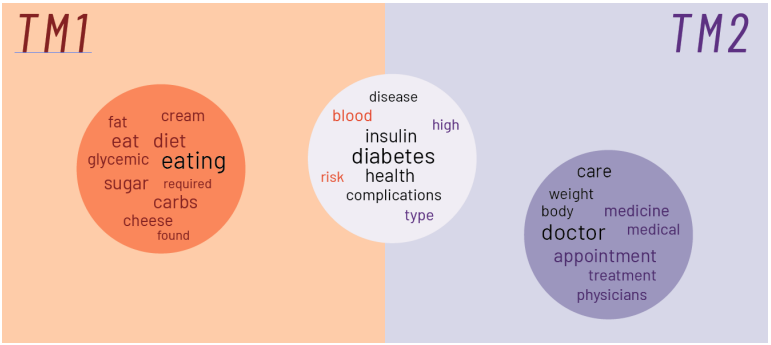


Figure 4.6: Topic coin clustering mockup. Each coin glyph represents one or more merged topics, with shared keywords in black and unique keywords in orange and purple corresponding to their respective topic model. Relative positions of each glyph encodes that topic's similarity to other topics as well as distinctiveness across the two models.

Initially, we explored a clustering topic coin approach similar to Oelke et. al. [34] by merging similar topics into coin glyphs and projecting topic distances in two-dimensional space (see the mockup in Figure 4.6). Within

each glyph, one or more merged topics would be shown in a tag cloud of topic keywords, with weight encoded by text size. Shared keywords from merged topics are shown in black, while words unique to one model's topic are coloured accordingly (orange for the left model, purple for the right model). The location of the topic glyph, as well as its background colour, would encode relative distinctiveness across models; i.e., a glyph farther to the left would represent a topic more unique to the left model. Relative coin distances would also encode topic similarity or segregation between pairs of topics.

As we considered this clustering paradigm, we extended this framework to address more comparison tasks. Particularly, the need to situate topics in the context of domain-specific knowledge led to the inclusion of ontology mappings in the interface. As well, we introduced the document view to address the unmet need to thoroughly explore document similarities and individual documents. Figure 4.7 shows the first mockups of these views during one of our brainstorming design sessions (Nov 2018, UBC).

Eventually, the coin glyph approach was replaced with a simpler parallel coordinates graph of topics, topic keywords, and topic alignment, in order to remove the additional problem of merging aligned topics as well as create more room for other visual encodings. Figure 4.8 shows a much more high fidelity mockup for topic alignment matrices, document-topic matrices, and the view for exploring individual documents.

For the document view, we initially encoded keyword strength with colour with bilinear interpolation between the two selected topics (see Figure 4.9). These colour gradients were deemed to be too difficult to perceive, and we experimented with varying interpolation, changing colours, and increasing the bin size of keyword weights to reduce the colour possibilities. In the final solution, this was simplified further to three static colours (blue and red for keywords in one model topic, and green for shared keywords), as seen in Figure 4.1 (bottom-center).

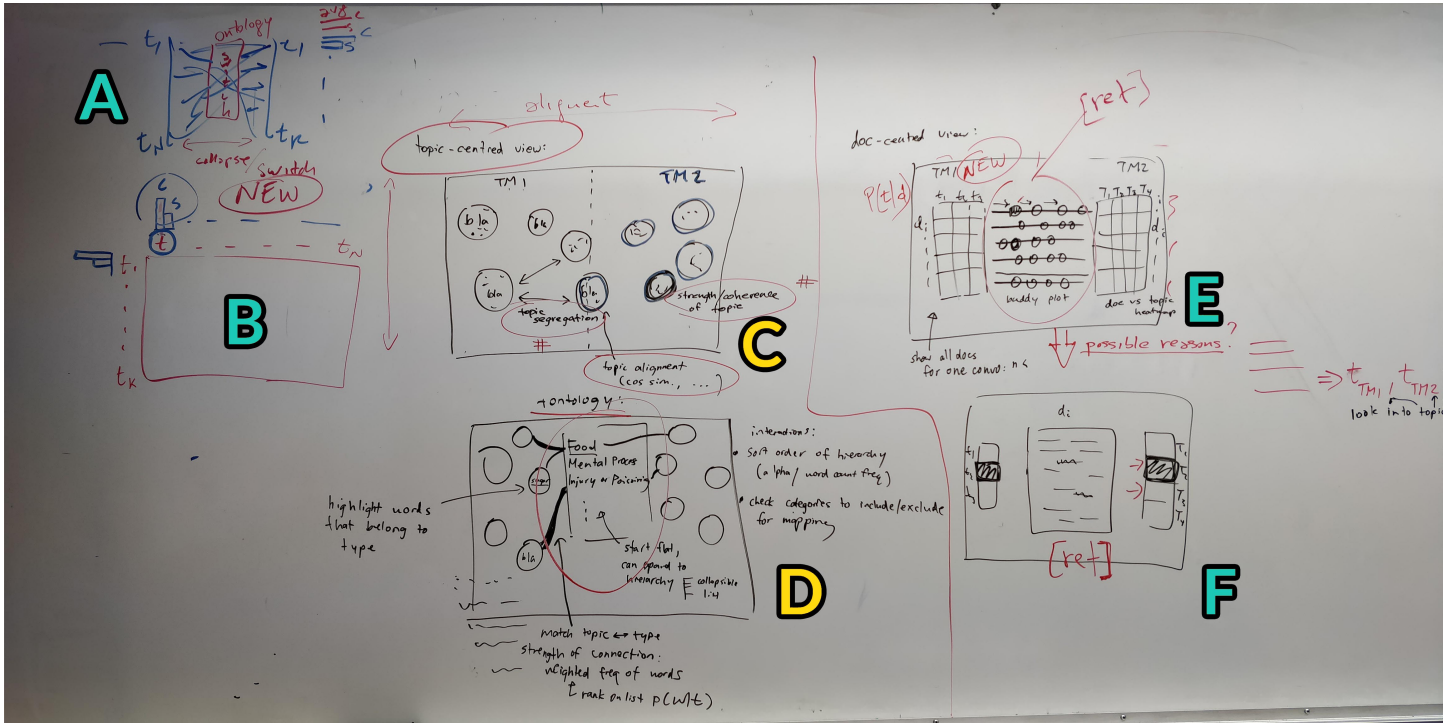


Figure 4.7: Whiteboard mockup of topic-centred and document-centred views, snapshot during brainstorming design session before finalizing the design of *OCTVis* (**adopted designs**, **discarded/modified designs**). (A) parallel coordinates graph to display ontology mappings; (B) topic alignment matrix; (C) clustering of topic coins; (D) clustering with ontology mapping; (E) document-topic matrices with buddy plots; (F) individual document view

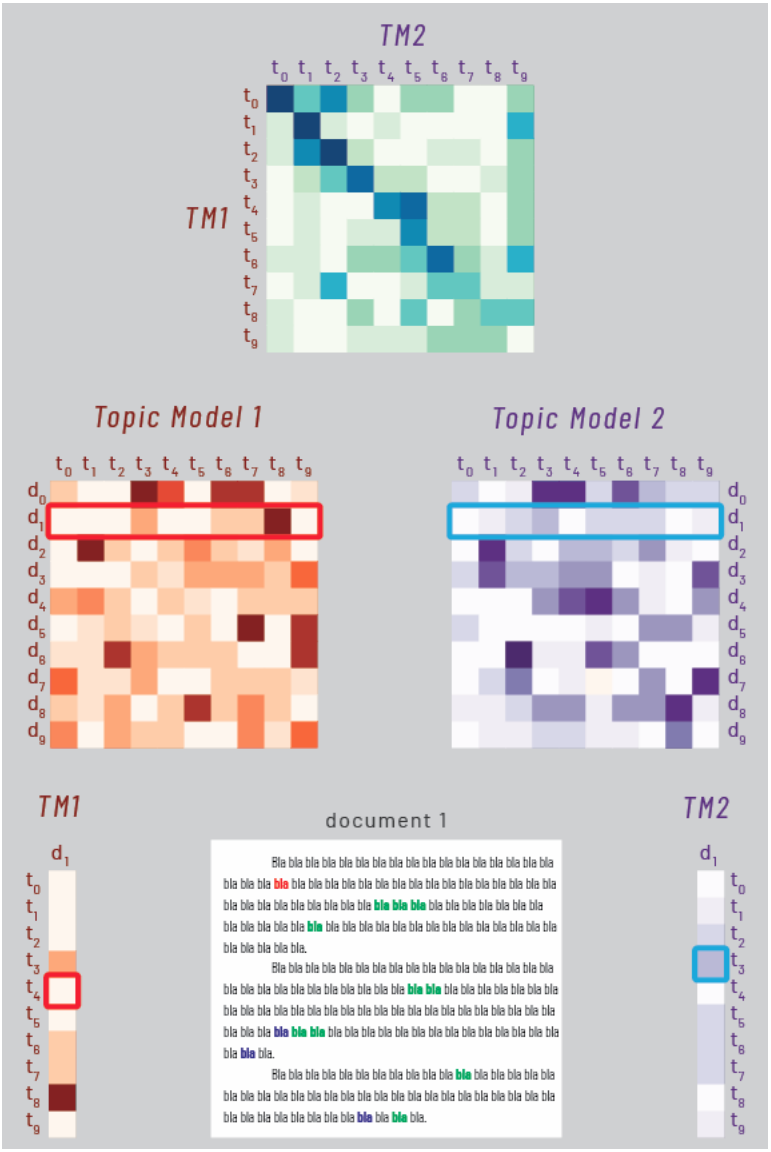


Figure 4.8: Mockup of topic alignment matrices, document-topic matrices, and document view.

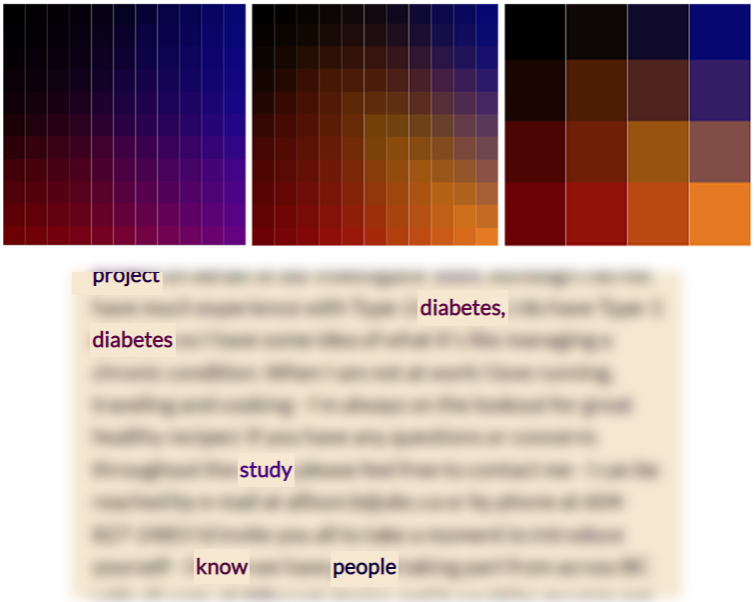


Figure 4.9: Colour binning for document keywords. *Top left*: bilinear interpolation between black and red (vertical, model A keyword weight), and black and blue (horizontal, model B weight). *Top centre*: bilinear interpolation along diagonals, bottom-left to top-right diagonal blending between model A (red) to shared (orange) to model B (blue), top-left to bottom-right diagonal blending with black based on keyword strength. *Top right*: same as centre, reduced to 4×4 colour bins. *Bottom*: document view with keywords coloured using top-left binning for topic weights, showing the difficulty in perceiving relative topic weight.

4.5 Implementation

Topic modeling and ontology mappings were computed with Python's `nltk` and `sklearn` libraries. In this thesis, we used Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) with 5 and 10 topics each. In order to annotate topic modeling results with ontology concepts, we use MetaMap [2], a toolkit that automatically codes text with the UMLS semantic concepts.

The *OCTVis* interface is a web app built with HTML, CSS, and Javascript. We used Sass[2] for programmable stylesheets and the D3 (v5) Javascript library[3] for data-driven visualization. We make abundant use of ES6 improvements to Javascript, including `async`/`await` for asynchronous file loading, `Promise.all` for parallel asynchronous calls, and arrow functions for D3 callbacks[4]. All source code can be found on `http://www.github.com/humfuzz/octvis`. An online demo of the interface can be found on `http://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/octvis/`[5].

We use the following input to *OCTVis*:

1. Raw text for each document in the corpus,
2. Comma-separated values for document-topic distributions from the results of two topic models, where each row is a document, each column is a topic, and each field is the weight of that topic for that document,
3. A list of unique vocabulary words in the corpus, separated by new lines,
4. Comma-separated values for word-topic distributions from the results of two topic models, where each row is a topic, each column is a word in the order of the vocabulary list, and each field is the weight of that topic for that word, and
5. Ontology mappings from each ontology term to vocabulary words, represented with a JSON `Object`.

These files are parsed and processed by `load.js`, which computes topic alignment, document similarities, per-topic metrics (e.g. segregation), and

[2] `https://sass-lang.com/`
[3] `https://d3js.org/`
[4] e.g. `(d, i, nodes)=>{... nodes[i]}` for binding `this`
[5] Documents used in case studies are redacted for ethics purposes.

ontology weights for each topic (see Chapter 3.1 for details). All data is represented in Javascript `Object`s and `Array`s.

`main.js` renders the *OCTVis* visualization in a webpage using SVG. Rectangular heatmaps and buddy plots use the standard D3 data join model to associate SVG selections with Javascript data. We take advantage of D3's force-body simulations to generate node-link graphs for topic alignment and ontology mapping, fixing nodes to computed positions and pausing the simulation on the first timestep.

The document view is a `foreignObject` `div` nested inside the SVG, and contains `span`s for each word in the document. Each word `span` is dynamically classed based on the selected topics to highlight keywords within that topic.

Facet locations are dynamically adjusted for the number of topics and documents in the data, and custom tweaks per dataset can be manually specified in metadata parameters. For instance, the top n keywords (default: 1% of the vocabulary size) per topic are highlighted in the document view, and n can be manually adjusted for each dataset.

Chapter 5

Case Studies

5.1 Methodology

To evaluate *OCTVis* on real data, we ran case studies with domain experts on two corpora of comments from online discussion forums: one containing comments from diabetes patients, the other containing comments from ovarian cancer patients.

Table 5.1: Case study corpora statistics.

		diabetes discussions	**ovarian cancer discussions**
	# comments	201	56,537
# words per comment	min	1	1
	max	524	5,577
	mean	110	99
	st. dev.	93	130

The topic models were built on corpora with very different sizes (see Table 5.1). For each corpus, the two topic models being compared were generated with differing methods: one with Latent Dirichlet Allocation (LDA) [3], and the other with Non-negative Matrix Factorization (NMF) [28], both set to generate five topics. For the ontology mapping we used the Unified Medical Language System (UMLS) Semantic Network [31], which describes 127 terms[6] ("concepts" or "semantic types") and 54 relations in the medical lexicon. We manually excluded 24 of these concepts that were too general or uninformative (for instance, *Qualitative Concept*, *Functional Concept*, and *Group Attribute*) and weighted the mapping from each concept to each word in the corpus using Equation 3.1. For each topic, we computed its segregation from the other topics in the same model using the average cosine similarity of pairwise topic vectors. As for the set of documents that were

[6]We used UMLS semantic network *2018AA*, the latest version at the time of writing.

actually shown in the visualization, in order to keep the case studies manageable in term of time, a sample of the comments was selected and shown: 34 comments for the diabetes corpus and 27 for the ovarian cancer corpus. We assume, given more time, the domain and NLP experts would consider different and possibly larger samples of the data to perform a more in-depth comparison of the topic models.

Our four domain experts were all clinicians that work with patients directly or facilitate online patient discussions. At the start of the case study, we provided the experts with a definition of topic modeling and ontologies, and asked about whether they thought topic modeling can be informative in understanding a possibly large set of documents, whether they had a clear idea what a 'good' topic model should look like, and whether or not ontologies are useful in evaluating topic model quality.

After this questionnaire, each domain expert was guided through the features of *OCTVis* on a fictional sample dataset and topics, and asked to demonstrate their understanding by answering questions based on performing sample tasks. If they could not answer those questions or anything else was unclear, more training was provided.

During the main study, they were presented with one of two datasets and were free to explore the data using the interface. Two of the experts had more time and were able to also explore the same corpus with two additional models comprising ten topics instead of five, and one of these experts was able to give feedback on both corpora as well. Since the goal of our interface involves facilitating communication between domain experts and NLP researchers, one of our NLP researchers discussed the results of the topic models and prompted topic model comparison with the domain expert throughout the main study. We were interested in how the interface, through topic model comparison with and without mapping to an ontology, could improve this communication and facilitate expert evaluation of the topic models. After the study, we asked the domain experts the same questions at the start to see whether their beliefs about topic models and ontologies changed, and asked for generic feedback on the interface.

We made a conscious decision to exclude the parallel buddy plot facet from the case study, as we felt that the visual encodings of the parallel buddy plots had a steep learning period relative to marginal benefits of the document distance information in the medical conversational corpora we considered. Since each document was a comment from discussion fora, they were relatively short (around 20-500 words each) compared to other classes of documents in NLP (for instance, a newspaper article or a journal paper); this meant that relative document distances were drastically different based

on small changes in the topics, and so had very little similarities between the two topic models. Without any significant patterns in the buddy plot data to compare document distances across models, we decided not to include this view in the case study. For corpora with longer documents, for instance, a collection of newspaper articles, the buddy plot view would likely be more informative. Testing our interface on such documents is left as future work.

5.2 Results

The case studies indicate that *OCTVis* was useful in comparing the two models. From the pre-study and post-study questionnaires (see Table 5.2), we found that all the experts reported that their idea of what a 'good' topic model should be became clearer after using the interface. Their belief that topic modeling can be informative in understanding a large set of documents was strong before and remained strong after the study. However, two experts lowered their assessment for the usefulness of ontologies, mainly because they found that the provided ontology was too generic.

Table 5.2: Pre-study and post-study questionnaire responses. Domain experts were asked how strongly they agreed or disagreed with the following statements (on a scale from 1-5, with 0.5 increments, where 1 is strongly disagree and 5 is strongly agree): (S1) Topic modeling can be informative in understanding a possibly large set of documents. (S2) I have a clear idea of what a 'good' topic model should look like. (S3) Ontologies are useful in evaluating topic model quality. For each expert and question, bold numbers indicate stronger or equal agreement *before* or *after* the study.

domain expert	S1		S2		S3	
	pre	post	pre	post	pre	post
e1	**4**	**4**	4	**4.5**	4	**4.5**
e2	4	**5**	3	**5**	**4**	3
e3	**4**	**4**	3	**4**	**4**	**4**
e4	**5**	**5**	4	**5**	**4**	2

Topic Alignment

All the domain experts used the topic alignment graph and heatmap to compare aligned topics and identify differences between a pair of similar topics.

Generally, they found that the NMF model generated more coherent topics than the LDA model. For the diabetes corpus, one domain expert found that the LDA model was more high level and the NMF model was more detailed. In the ovarian cancer discussions, another expert found that the LDA topics were more medical focused while the NMF topics were more psychosocial. This led the domain and NLP experts to consider the possibility of combining the two models in future work. For instance, speculatively, topics from NMF could be used to refine topics form LDA into more fine-grained subtopics. Alternatively, topics from NMF and LDA could be joined in a single model with broader semantic coverage. In terms of topic segregation, the per-topic segregation metric was found to be relatively uninformative across our case studies, and one expert expressed the desire to see more quantitative metrics across both topic models, such as the fraction of the concepts mapped in the ontology.

The two domain experts that were able to see the results of ten-topic topic models in addition to five-topic models found that the interface scaled well for the increased visual information. One expert initially found using the alignment matrix heatmap less intuitive than the node-link graph to compare topic models with five topics, but with ten topics the node-link graph became more cluttered, and the alignment matrix heatmap was much more effective for identifying pairs of strongly aligned topics, an indication that redundantly displaying such information can be beneficial. Within the ten-topic models, both the experts identified within-model topics that were very similar in both keywords and document distributions and suggested that merging these topics would result in more cohesive topics.

Ontology Mapping

While two of our experts found the ontology mapping to be a useful high-level overview of the topics discussed, the other two found that the mapping as-is still contained many concepts that were too general or uninformative. Those that found the ontology mapping useful were able to re-evaluate their understanding of topics based on their ontology mappings. In the ovarian cancer corpus, one topic with top keywords "chemo just time taxol like did good" mapped strongly to the ontology concept "Therapeutic or Preventive Procedure", and no other topics in either model mapped as strongly to this concept. With this information, the two experts that explored this corpus were able to identify this topic as a cohesive and well-segregated topic about cancer treatment that the other model failed to capture.

All our domain experts provided useful feedback about how to enhance

the utility of the ontology mapping. For instance, one clinician was unclear about what the ontology term labeled "Finding" represented. Even after we clarified that it referred to *medical* findings, they believed that it was too general and might be better broken down into subcategories including "test results" or "symptoms". All the experts expressed a need to dynamically alter or explore the ontology concepts shown, through filtering, searching, or manually adding in categories of interest. Particularly, when the topic models missed important and expected topics including "medication", the ontology view could show this gap in coverage. Furthermore, it was suggested that we could use these gaps to seed future topic models to improve the modeling results.

Exploring Documents

Two of the clinicians expressed little interest in exploring individual documents, as they were more concerned about the overall topics discussed by the patients, while others used in-depth exploration of the documents to assess topic quality and improve their understanding of the topics. For the diabetes corpus, the document-topic distribution and highlighted keywords in individual documents helped confirm labels for topics including "exercise" and "diet", as well as determining that a "medication" topic was missed by both models. This was less the case for the ovarian cancer corpus, likely because the corpus used to generate the topic models was much larger than the sample documents shown in the interface. The domain experts that compared topics at the document keyword level were often able to qualify similar topics with more nuance - for example, while comparing very similar topics relating to "surgery" in the ovarian cancer corpus, one model's topic had more cohesive keywords (less irrelevant keywords) in a document than the other model and was perceived to capture the topic in the text more precisely.

Chapter 6

Discussion and Future Work

Overall, our case studies showed that *OCTVis* enabled domain experts to effectively compare topic model results and suggest topic modeling improvements to NLP experts. There was no clear trend in valuing broad comparisons against an ontology over fine-grained comparisons against individual documents, and our evaluation suggests that exploring both contexts enhances the evaluation of topic model quality and allows for more subtle comparisons. The ontology mapping acts as a high-level overview of the corpus to give the two topic models a backbone to compare against, while exploring documents in-depth enables detailed validation of topic coherence and coverage.

As mentioned in Chapter 5.1, we excluded the parallel buddy plots from our case studies due to the limited document distance information they provided. Given the short documents in our corpora, relative document distances had little similarity across the two topic models, and the buddy plots had no significant patterns to identify unique documents in the distance space. We hope to test the effectiveness of this encoding on corpora with longer documents, where we would expect to see more similarities in document distances between differing topic models.

Given our visual framework, a future application exists for comparing ontology-based topic model results on differing corpora, similar to the comparative framework set by Oelke et. al. [34]. For example, clinicians may be interested in how discussions differ for patients from one region to another, and our interface could be adapted to allow for this comparison. As this involves comparison of corpora rather than topic models, a future task model and subsequent implementation would need to reflect this change.

Existing work [23, 25, 26] has found human-in-the-loop visual analytics to be extremely beneficial for exploring conversational text data, especially asynchronous conversations including online blogs and discussions much like the corpora in our case studies. A future extension to our comparison framework involves displaying the conversational structure of these corpora alongside our ontology-based comparison tools. Additionally, conversational text is well suited for hierarchical topic modeling techniques

[13, 25] that allow users to view high-level topics and drill down for more detailed subtopics, and these hierarchical topic models would be invaluable for comparison tasks. Both conversational text analytics and hierarchical topic modeling methods would increase the scalability of our solution for corpora with a large number of documents by aggregating topic modeling information on collapsible conversational documents.

These human-in-the-loop techniques often involve interactive topic model revision [15, 16, 25], and our case studies have shown the need for more interactive tools in our comparative framework. Techniques for interactive topic modeling, including seeding the topic model with keywords and merging/splitting/deleting topics, are all tools our domain experts expressed a desire to use, particularly when applied to the ontology space. Allowing domain experts to build more relevant ontologies and enhance the existing ontology network would enhance the effectiveness of ontology-based comparison of topic models. Similar to work in building knowledge graphs [21, 36], we can use this to have custom ontologies built for specific domain needs.

With larger numbers of topics, ontology concepts, and documents, the scalability of our interface becomes a concern. Aforementioned interactive techniques including filtering and drilling down can tackle the issue of scale on many data dimensions. Topics, documents, and ontologies can be clustered hierarchically and allow users to explore the data space at multiple granularities. For instance, for large ontologies, analysis could focus on more fine-grained subtrees that are specific for relevant topics (e.g. cancer or diabetes sub-ontologies for our corpora) and allow interactive expansion and collapse of these concepts. While the alignment encoding with parallel coordinates can become cluttered with dozens of topics or concepts, heatmap matrices and other space-filling idioms can scale more effectively; our topic-ontology mappings can be redundantly encoded with these visual encodings as well.

In terms of the algorithmic choices that were made on how the data model is derived, several future alternatives could be considered. For instance, in order to compute topic alignment, we used the generic cosine similarity measure (see Equation 3.2). If, for instance, the two topic models involved were *both* probabilistic, distances that are specific to probability distributions, like KL-divergence, could be more appropriate [40]. Furthermore, in computing the mapping from topics to ontology concepts, we have assumed that the corpus is deterministically labeled with ontology concepts. However, this might not be the case for the given ontology, for which the labeling could be probabilistic (i.e., a word in the corpus is assigned to different ontology concepts with different probabilities). When this happens,

Equation 3.1 should be revised accordingly:

$$weight(t_k, c_i) = \sum_{w_j \in V,\, c_i \in C} \boldsymbol{P(c_i|w_j)} P(w_j|t_k), \tag{6.1}$$

where we include the probability $P(c_i|w_j)$ of a concept c_i given a word w_j, and sum over all words w_j in the corpus V and all concepts c_i in the ontology C.

Chapter 7

Conclusion

Our visual analytics approach to ontology-based comparison of topic models is designed to inform NLP experts and domain experts about the quality of alternative topic modeling methods situated within the context of a domain ontology. We expand on prior work in topic modeling visualization by enabling topic comparison at multiple granularities, from high-level domain-specific mappings to ontology concepts to in-depth document keyword exploration. To the best of our knowledge, we are the first to perform an evaluation of a topic modeling comparison interface with potential users. Such evaluation, through case studies with clinicians, has shown a clear improvement in understanding topic modeling, enabling communication between domain experts and NLP experts to evaluate and improve novel topic modeling techniques. We hope to extend our work in the future by facilitating comparison of different corpora, incorporating more interactive techniques for human-in-the-loop analysis, and enhancing comparison of hierarchical topic models and conversational text.

Bibliography

[1] E. Alexander and M. Gleicher. Task-driven comparison of topic models. *Trans. Visualization and Computer Graphics (TVCG)*, 22(1):320–329, Jan 2016. ↑1, ↑6, ↑7, ↑17, ↑18, ↑20, ↑22, ↑24, ↑29, ↑30

[2] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal American Medical Informatics Assoc.*, 17(3):229–236, 2010. ↑17, ↑36

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal Machine Learning Research*, 3:993–1022, Jan 2003. ↑15, ↑38

[4] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proc. German Society Comp. Ling. & Language Tech. (GSCL)*, pages 31–40, 2009. ↑18

[5] N. Cao, D. Gotz, J. Sun, Y. Lin, and H. Qu. SolarMap: Multifaceted visual analytics for topic exploration. In *IEEE 11th Intl. Conf. Data Mining*, pages 101–110, Dec 2011. ↑3

[6] Giuseppe Carenini, Gabriel Murray, and Raymond Ng. Methods for mining and summarizing text conversations. *Synthesis Lectures on Data Management*, 3(3):1–130, 2011. ↑1

[7] K. Chan, X. Lou, T. Karaletsos, C. Crosbie, S. Gardos, D. Artz, and G. Ratsch. An empirical analysis of topic modeling for mining cancer clinical notes. In *IEEE 13th Intl. Conf. on Data Mining Workshops (ICDMW)*, pages 56–63, Los Alamitos, CA, USA, dec 2013. IEEE Computer Society. ↑1

[8] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 2009. ↑1, ↑9

[9] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *Trans. Visualization and Computer Graphics (TVCG)*, 19(12):1992–2001, Dec 2013. ↑3

[10] Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proc. Intl. Conf. Machine Learning (ICML)*, 2013. ↑1, ↑8, ↑9, ↑10

[11] E. Talley M. Leenders D. Mimno, H. M. Wallach and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272. Association for Computational Linguistics, 2011. ↑18

[12] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *IEEE Conf. Visual Analytics Science and Technology (VAST)*, pages 231–240, Oct 2011. ↑3, ↑4

[13] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *Trans. Visualization and Computer Graphics (TVCG)*, 19(12):2002–2011, Dec 2013. ↑4, ↑6, ↑44

[14] Marek Dudáš, Steffen Lohmann, Vojtch Svátek, and Dmitry Pavlov. Ontology visualization methods and tools: a survey of the state of the art. *The Knowledge Engineering Review*, 33, 07 2018. ↑10, ↑11, ↑12

[15] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *Trans. Visualization and Computer Graphics (TVCG)*, 24(1):382–391, Jan 2018. ↑9, ↑10, ↑31, ↑44

[16] M. El-Assady, F. Sperrle, O. Deussen, D. Keim, and C. Collins. Visual analytics for topic model optimization based on user-steerable speculative execution. *Trans. Visualization and Computer Graphics (TVCG)*, 25(1):374–384, Jan 2019. ↑6, ↑9, ↑18, ↑44

[17] Sean Falconer, R Ian Bull, Lars Grammel, and Margaret-Anne D. Storey. Creating visualizations through ontology mapping. pages 688–693, 03 2009. ↑10

[18] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan. ThemeDelta: Dynamic segmentations over temporal topic models. *Trans. Visualization and Computer Graphics (TVCG)*, 21(5):672–685, May 2015. ↑4, ↑7

[19] M. Gleicher. Considerations for visualizing comparison. *Trans. Visualization and Computer Graphics (TVCG)*, 24(1):413–423, Jan 2018. ↑18

[20] M. Glueck, M. P. Naeini, F. Doshi-Velez, F. Chevalier, A. Khan, D. Wigdor, and M. Brudno. PhenoLines: Phenotype comparison visualizations for disease subtyping via topic models. *Trans. Visualization and Computer Graphics (TVCG)*, 24(1):371–381, Jan 2018. ↑10, ↑13

[21] Udo Hahn, Martin Romacker, and Stefan Schulz. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. In *Biocomputing 2002*, pages 338–349. World Scientific, 2001. ↑44

[22] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 115–123, Oct 2000. ↑4, ↑7

[23] E. Hoque and G. Carenini. ConVis: A visual text analytic system for exploring blog conversations. In *Proc. 16th Eurographics Conf. Visualization (EuroVis)*, pages 221–230. Eurographics Association, 2014. ↑3, ↑43

[24] Enamul Hoque and Giuseppe Carenini. MultiConVis: A visual text analytics system for exploring a collection of online conversations. In *Proc. 21st Intl. Conf. Intelligent User Interfaces (IUI)*, pages 96–107, New York, NY, USA, 2016. ACM. ↑5, ↑6

[25] Enamul Hoque and Giuseppe Carenini. Interactive topic hierarchy revision for exploring a collection of online conversations. *Information Visualization*, 2018. ↑6, ↑43, ↑44

[26] Indratmo, Julita Vassileva, and Carl Gutwin. Exploring blog archives with interactive visualization. In *Proceedings of the Working Conf. on Advanced Visual Interfaces*, AVI '08, pages 39–46, New York, NY, USA, 2008. ACM. ↑43

[27] Akrivi Katifori, Constantin Halatsis, George Lepouras, Costas Vassilakis, and Eugenia Giannopoulou. Ontology visualization methods - a survey. *ACM Computing Surveys*, 39:10, Nov 2007. ↑10

[28] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999. ↑38

[29] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, 2012. ↑3

[30] Shixia Liu, Michelle Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. TIARA: Interactive, topic-based visual text summarization and analysis. volume 3, pages 543–552, 01 2009. ↑4, ↑7

[31] Alexa T McCray. An upper-level ontology for the biomedical domain. *International Journal of Genomics*, 4(1):80–84, 2003. ↑15, ↑38

[32] T. Munzner. A nested model for visualization design and validation. *Trans. Visualization and Computer Graphics (TVCG)*, 15(6):921–928, Nov 2009. ↑14

[33] Tamara Munzner. *Visualization analysis and design*. AK Peters Visualization Series, CRC Press, 2014. ↑1, ↑15

[34] D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen. Comparative exploration of document collections: a visual analytics approach. *Computer Graphics Forum*, 33(3):201–210, Jul 2014. ↑1, ↑7, ↑8, ↑31, ↑43

[35] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume, and Lise Getoor. Understanding MOOC discussion forums using seeded LDA. In *Proc. 9th Workshop Innovative Use of NLP for Building Educational Applications*, pages 28–33. Association for Computational Linguistics (ACL), Jun 2014. ↑1

[36] David Sánchez and Antonio Moreno. Learning non-taxonomic relationships from web documents for domain ontology construction. *Data & Knowledge Engineering*, 64(3):600–623, 2008. ↑44

[37] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer. On the beauty and usability of tag clouds. In *12th Intl. Conf. Information Visualisation*, pages 17–25, Jul 2008. ↑3, ↑24

[38] Arjun Srinivasan, Matthew Brehmer, Bongshin Lee, and Steven M. Drucker. What's the difference?: Evaluating variations of multi-series bar charts for visual comparison tasks. In *Proc. CHI*, pages 304:1–304:12, New York, USA, 2018. ACM. ↑7

[39] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 952–961, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. ↑1

[40] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007. ↑44

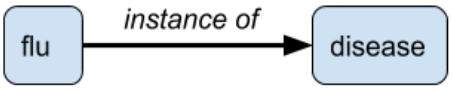
[41] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo. TopicPanorama: A full picture of relevant topics. *Trans. Visualization and Computer Graphics (TVCG)*, 22(12):2508–2521, Dec 2016. ↑3, ↑5, ↑6

Appendix A

Pre-Study Questionnaire

Topic modeling is the process of extracting the topics covered by a set of documents, or corpus. For instance, a set of documents can cover two topics: animals and food.

An ontology is a set of concepts and relationships describing knowledge in a particular domain. For instance, in the medical domain, two concepts could be "flu" and "disease"; they are related by an instance of relation: flu is an instance of disease.



Please rate how strongly you agree or disagree with each of the following statements with respect to topic modeling.

Topic modeling can be informative in understanding a possibly large set of documents.

strongly disagree	disagree	neutral	agree	strongly agree	don't know

I have a clear idea of what a good topic model should look like.

strongly disagree	disagree	neutral	agree	strongly agree	don't know

Ontologies are useful in evaluating topic model quality.

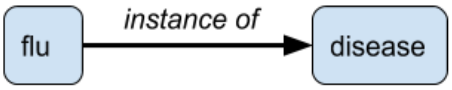
strongly disagree	disagree	neutral	agree	strongly agree	don't know

Appendix B

Post-Study Questionnaire

Topic modeling is the process of extracting the topics covered by a set of documents, or corpus. For instance, a set of documents can cover two topics: animals and food.

An ontology is a set of concepts and relationships describing knowledge in a particular domain. For instance, in the medical domain, two concepts could be "flu" and "disease"; they are related by an instance of relation: flu is an instance of disease.



Please rate how strongly you agree or disagree with each of the following statements with respect to topic modeling.

Topic modeling can be informative in understanding a possibly large set of documents.

strongly disagree	disagree	neutral	agree	strongly agree	don't know

I have a clear idea of what a good topic model should look like.

strongly disagree	disagree	neutral	agree	strongly agree	don't know

Ontologies are useful in evaluating topic model quality.

strongly disagree	disagree	neutral	agree	strongly agree	don't know

Please answer the following questions.

What did you like about the interface? What did you find to be useful?

What did you dislike about the interface? How can the visualization be improved?

What did I learn about the dataset, thanks to the topic models and ontology?

Which topic model better represents the text?
Please justify your findings.

How can either topic model be improved?

Appendix C

Case Study Sample Dataset

[document 1]
I love animals. Animals are so great. I love cute pandas and giraffes and dogs and horses and cats and chimpanzees.

[document 2]
Me too! Animals are so cute.

[document 3]
Here is my recipe for banana bread. Makes one delicious loaf.
2 bananas
1 cup sugar
1 egg, beaten
1 cup flour

1. Preheat oven to 350F.
2. Mash bananas and add egg, sugar, and flour
3. Pour batter into loaf pan. Bake for 1h.
4. Remove from oven and serve.

[document 4]
I love feeding my cats banana bread, they gobble it up!

Appendix D

Case Study Script

D.1 Introduction and pre-study questionnaire

Today we'll be investigating comparison of topic models, particularly those seeded with existing ontologies, with the goal of assisting NLP researchers develop novel techniques as well as improving communication between NLP researchers and domain experts when evaluating new methods.

Topic modeling is the process of extracting the topics covered by a set of documents, or corpus. For instance, a set of documents can cover two topics: animals and food.

An ontology is a set of concepts and relationships describing knowledge in a particular domain. For instance, in the medical domain, two concepts could be "flu" and "disease"; "disease" would have a superclass relationship with "flu".

[Show example image in pre-study questionnaire.]

Thank you for participating in our study. The whole process today will last about one hour. My name is Amon and Im the designer of the interface, and Hyeju is the natural language processing (or NLP) researcher who will help facilitate this case study with you, a domain expert.

Do you have any questions about these?

Do you have any questions before we get started?

Okay, please feel free to ask questions anytime.

First you will answer a short pre-study questionnaire. Then, I will walk you through the interface with some sample data. After that, we will move to the main portion of the study, which will involve you exploring some of your data and comparing topic models with Hyeju. At the end of the study, Ill ask a few more questions to get your feedback.

Before we start, please answer a few questions.

[Have domain expert fill the pre-study.]

D.2 User training

Thank you, now we will demonstrate the features of the topic comparison interface. Here is a sample corpus, a dataset of four documents.

[Hand paper with 4 sample documents printed out.]

Now I will show you the results of two topic models on this dataset.

The visual interface presents two topic models side-by-side - the left model, model A, is red, and the right model, model B, is blue. For instance, one of the red topics is about the words "animals cute love...", and one of the red topics is about the words "eggs animals delicious...". Do you have any questions about this?

Above, the **Topic Overview** presents the top keywords of each topic, as well as per-topic metrics to estimate different aspects of topic quality. Right now, they are showing topic segregation - the longer the bar, the more distinct that topic is from the other topics within that model.

[Show/hide computed per-topic metrics.]

Do you have any questions about this?

The **Topic Alignment** facet, in green, consists of a graph between the topics of both models and a heatmap. Both these components indicate the similarity between pairs of topics across models with the "brightness" of the green colour.

[Highlight alignment heatmap cells to show corresponding linked topics.]

Do you have any questions about this?

[Toggle the Topic Alignment and Ontology Alignment facets.]

The orange **Ontology Alignment** facet replaces the Topic Alignment facet to show the strongest ontology concepts in the corpus, sorted in decreasing strength combined across both models. The strength of each alignment between ontology concept and topic is encoded by the colour and width of the line connecting the two. Do you have any questions about this?

Below each model's topics, the **Document Overview** shows the distribution of topics within each document. The colour of each cell indicates the strength of that topic (column) in that document (row). In the bottom centre of the interface, the **Document View** displays the text of a single document, as well as the top keywords between a selected topic from each model. Red and blue colours indicate keywords belonging to a single models selected topic, and green indicates a keyword shared across both topics.

[Click document row in Document Overview to view that document.]

[Click topic cells in the document heatmap to view keywords for that topic within that document.]

Do you have any questions about this?

Now, we will ask you a few questions about the dataset to check that you understand how to navigate the interface. At any time, you can ask questions and clarify your understanding with either of us.

[Give the domain expert control, and clarify if they don't know how to do any task.]

Can you identify the most similar topics across the two models?

In document 3, what are the two strongest topics in model A and model B?

Between these two topics, what keywords are shared within this document?

In model B, which topic is most prevalent throughout the corpus?

If you toggle the Ontology view, which concept is topic B3 most strongly mapped to?

Which concept is topic A2 most weakly mapped, but not zero?

By examining the ontology mappings, the documents, and the topics, which topic model would you pick as the better topic model to represent the data?

D.3 Main study

Now we will begin the study. You are going to be comparing two topic models on a conversation from an online discussion forum for [diabetes/ovarian cancer] patients mapped against a medical ontology, the Unified Medical Language System. This will be an open ended exploration between you and the NLP researcher. Imagine that Hyeju has come to you with these two topic models, and needs your domain knowledge to help evaluate and compare the results of these two models to improve future models and summarize the topics being discussed. Feel free to ask questions and communicate with each other, and please think out loud so that I can understand your experience.

D.4 Post-study questionnaire and debrief

Please answer the following post-study questionnaire.

[Have domain expert fill the post-study.]

Thank you very much again for your participation. Do you have any other comments or questions?