Knowledge engineering on PhD stories: bottom-up vs. goal-directed approach

VIET BACH NGUYEN[1], STANISLAV KRUML[1], VOJTĚCH SVÁTEK[1], ÓSCAR CORCHO[2], MAREK DUDÁŠ[1] and GOLLAM RABBY[1]

[1] *Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Praha 3, Czech Republic*
E-mail: viet.nguyen@vse.cz krus02@vse.cz svatek@vse.cz marek.dudas@vse.cz rabg00@vse.cz
[2] *Universidad Politécnica de Madrid, Madrid, Spain*
E-mail: ocorcho@fi.upm.es

Abstract

Stories of concrete PhD students, reflected in various information sources, are potentially a useful source of reflections and analytics on the underlying generic processes and regulations, and can subsequently give rise to guidelines and hints making the life of future PhD students easier. The collected information on PhD stories is of quite heterogeneous nature, from structured 'master data' from study information systems, through various partial records in broader (e.g., publication) databases, to textually expressed chronologies, impressions, and lessons learned prepared by the PhD students themselves. We explored two fundamentally different approaches to arrive at a structured representation of individual PhD stories. The first approach is bottom-up, relying on aspect-based sentiment analysis (SA) as an established NLP method. Training data was collected in the form of 35 textual stories (or, more broadly, documents reflecting the PhD experience of the author), manually labeled by volunteers, and submitted to state-of-the-art SA tools. The second approach is primarily manual in its nature, though amenable to knowledge/database support: population of a knowledge graph (or, concept map) from primarily structured sources as well as insider information, guided by a hierarchical system of (PhD study) goals. The task was accomplished, as a mere proof of concept, by two senior academics with ample experience in supervising PhD students. Both approaches appear complementary, but insufficient overall, to date. While the bottom-up one can much better scale, the result is largely incomplete in its coverage. The goal-directed approach is, in turn, extremely demanding on human expertise and laborious due to lack of proper tooling. Based on the gained experience, we discuss conditions that could allow making knowledge engineering on PhD stories more efficient and effective in the future.

1 Introduction

The research follows up with an initial pilot Nguyen et al. (2021), which however only featured the clean bottom-up approach in the form of thematic codes being provisionally assigned to twelve documents by a single researcher, and several ad hoc variants of graph modeling, without the support of any explicit guidelines. The current research, in contrast, newly introduces aspect-based sentiment analysis as a particular computational technique, in the bottom-up approach, and standardizes the graph modeling (goal-driven) approach through common guidelines. The total size of the collection has doubled (tripled). Explicit formulation of requirements allowing for effective knowledge engineering in the given domain is yet another contribution of the paper.

The structure of this paper is as follows:

2 Related research

2.1 PhD story analysis

2.2 Academic knowledge graphs

2.3 Biography modeling

3 Bottom-up analysis

3.1 Data collection

The PhD story data used in this research are mostly acquired from websites where they are published by their respective authors or by the permitted publishers. Apart from that, we also have gathered several "offline" stories via face-to-face interviews. Fig. 1 briefly shows our data collection strategy along with the proportion and statistics of the gathered resources by each document type.

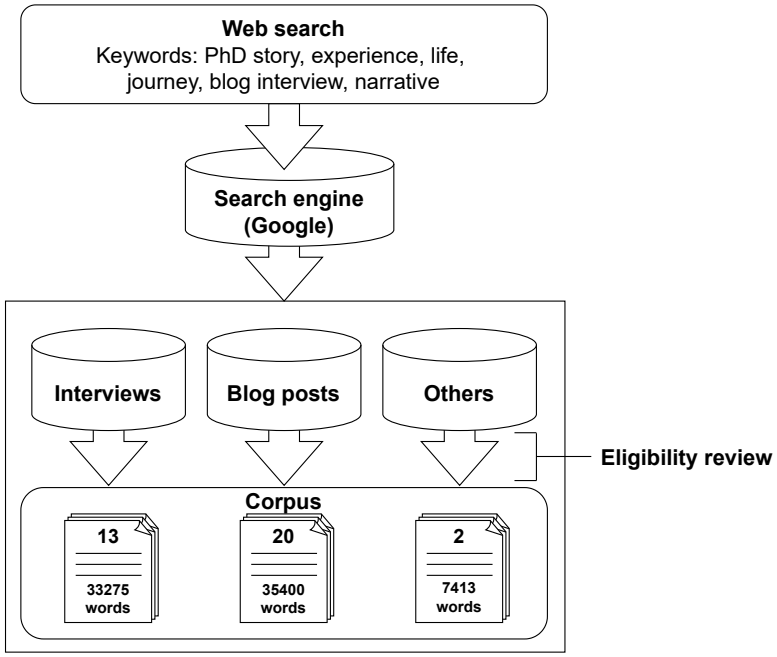

Figure 1 Data collection strategy

To collect stories that are published online, we followed a simple keyword-based search strategy on Google to produce results in the form of links. In this search strategy, suitable keywords are defined to search for PhD stories. These keywords are *PhD story*, *my PhD story*, *my PhD journey*, *PhD narrative*, *PhD experience*, *PhD blog*, *PhD interview*, *PhD life*, and some combinations. Next, we visited each link and quickly scanned the story content to determine whether a story is eligible for further analysis. Our selection criteria are not strictly defined, since every story is unique in writing style, form, and thoughts being conveyed. However, the most important front-line properties we looked for in the stories during the web search are language (only English), summarizing story-telling intent, sufficient story length (e.g., approximately 500 words or more), high coverage of multiple PhD life aspects (e.g., university life, well-being, private life, life after PhD), high density in retrospective narrations (e.g., lesson learned, experience, situations, decisions), chronological composition (e.g., sectioned into academic years or by aspects), and recentness of the story (e.g., we only consider modern PhD experiences, thus only stories published in the last time range of 20 years are collected, this should exclude PhD programs active in the 1990s and earlier).

In other words, the search results were filtered[1], and if the data do not satisfy the criteria mentioned above, they are omitted. In the case of "offline" interviews, we guided the interviewees with our open interview questions to allow them to speak freely but also to gather the needed information. These interview records (and 1 vlog) were later transcribed into text and included in the final dataset. Other interviews in the dataset were already in textual form.

Our dataset is a corpus of a total of 35 eligible PhD stories in textual form with about a total of 76 088 words, ranging from 500 words to 6 000 words each, the average word count is 2 174 and the median is 1 327. We classify the document types into blog posts (20), interviews (13), and others (1 vlog and 1 web article).

3.2 Document categorization

Since the nature of the documents varied, we manually skimmed them so as to establish categories (clusters) of documents having similar natures and/or destinations. ...

3.3 Aspect system design

We designed the aspects via brainstorming where we listed out different areas and topics that a PhD student may talk about. Also, with the consultancy of our senior members, we created the following tree of aspects to be used during aspect annotation.

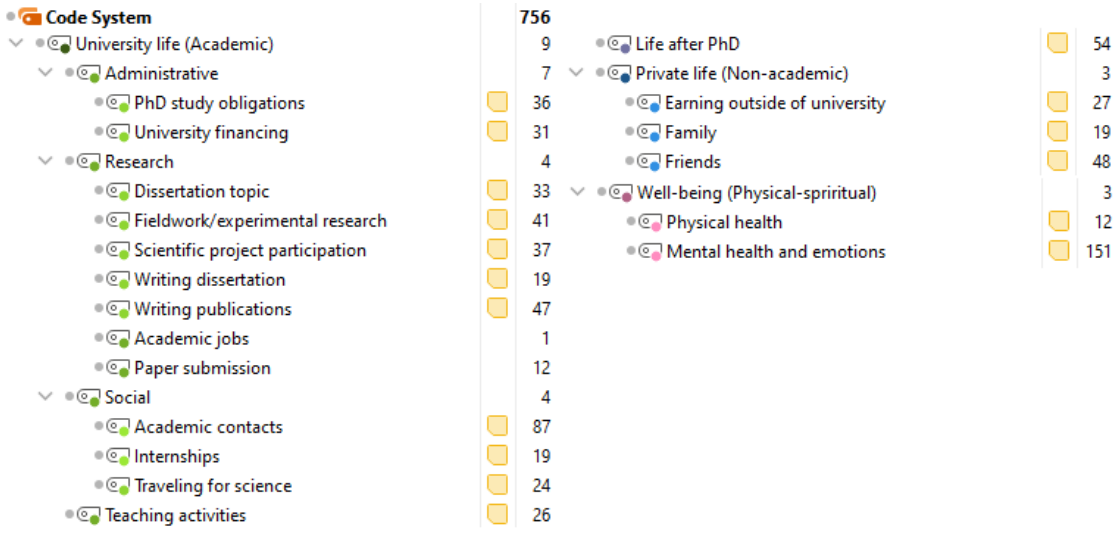

Figure 2 Annotation codes – aspects

3.4 Aspect annotation

We annotated the collection of PhD stories using MAXQDA software. For this annotation of aspects, we had a team of 6 people annotating 24 stories (at the time we only had 24 stories), see Tab. 1. Each team member was responsible for annotating 6 stories so that 12 stories are annotated once and the 12 stories are annotated twice by two people in hopes of measuring the annotators' agreement. We split the 24 stories into 6 pools evenly so that the time spent in each

[1]We do not filter the stories based on the research field of the authors, but we initially wanted to limit our search scope to Applied Informatics only. We decided to drop this filtering criterion end widen the scope because the initial number of search results was too few. We then expanded our search scope to the Computer Science domain, but again the search yielded an insufficient amount of results. Therefore, after removing this limitation, our final dataset contains stories from different research domains, such as biology, physics, maths, engineering, literature, etc.

story pool would be approximately the same as in other pools. The time measure estimation is based on the total word count of the stories in each pool. We approximated the reading and annotating speed to 100 words per minute (the average reading speed of an adult person is around 238 words per minute). The reported time spent in each pool by the annotators turned out to be approximately the same as estimated. In this aspect annotation phase, we have produced a total of 756 annotations.

pool	files to annotate	time estimate	words	annotator	annotations
A	09, 13, 17, 18, 20, 21	119 min.	12 182	Vojta	105
B	02, 06, 08, 13, 14, 19	114 min.	11 630	Standa	146
C	01, 03, 06, 08, 11, 15	120 min.	12 245	David	167
D	01, 04, 05, 10, 12, 23	110 min.	11 083	Marek	104
E	03, 16, 17, 19, 22, 23	123 min.	12 515	Bach	138
F	07, 10, 12, 15, 21, 24	122 min.	12 419	Gollam	96

Table 1 Aspect annotation pools

3.5 Polarity and emotion annotation

Aside from three-valued polarity, we also used this opportunity to acquire information on the *surprise* emotion. The reason is that the surprise emotion is one that does not inherently lean towards positive nor negative polarity, yet can be an important ingredient in a PhD student's story – whether in the context of the actual research (succeeding or failing against odds) or, e.g., social relationships within the lab or school.

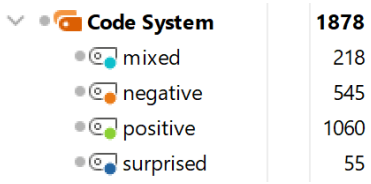

Figure 3 Annotation codes – sentiments

We also annotated our collection of PhD stories with sentiment values using MAXQDA software. This time we had a different team of 8 people manually annotate all 35 stories using 4 annotation codes which are *Positive*, *Negative*, *Mixed*, and the mentioned *Surprise* emotion. For this annotation task, we divided our corpus of stories into 8 distinct pools with 6–7 stories each so that 18 stories are annotated twice by two people, and the other half of the stories are annotated once. Each pool represents about 155 minutes of reading and annotating time of about 15 500 words (we stick with the speed of 100 words per minute from the aspect annotation phase). We calculate these estimated time values based on the word count of each document and distribute the stories into the pool evenly by estimated time spent, see Tab. 2. The reported time spent in each pool by the annotators turned out to also be approximately the same as estimated.

pool	files to annotate	time estimate	words	annotator	annotations
A	17, 20, 22, 25, 29, 31, 32	151 min.	15 105	Dominik	445
B	08, 13, 14, 15, 28, 34, 35	146 min.	14 642	Šimon	274
C	06, 08, 10, 19, 27, 30	160 min.	16 045	Michael	128
D	01, 02, 04, 05, 23, 25, 27	161 min.	16 080	Dominika	439
E	01, 03, 13, 17, 21, 23, 26	149 min.	14 863	Karel	239
F	07, 10, 12, 15, 24, 33, 35	151 min.	15 069	Maxim	116
G	03, 06, 09, 18, 19, 29	151 min.	15 148	Kateřina	124
H	11, 12, 16, 21, 31, 33	159 min.	15 876	Marek	113

Table 2 Sentiment annotation pools

Annotators were then randomly assigned to these pools. We also provided the annotators with a common guiding document with instructions explaining the necessary steps for the annotations. Our definitions of annotation values are the following:

- *Positive* – there is an explicit or implicit clue in the sentence suggesting the writer's attitude towards or judgment of the subject is positive (thankful, excited, optimistic, overjoyed, happy, inspired, etc.),
- *Negative* – there is an explicit or implicit clue in the sentence suggesting the writer's attitude towards or judgment of the subject is negative (critical, angry, disappointed, pessimistic, sarcastic, mocking, bored, complaint, etc.),
- *Mixed* – there is an explicit or implicit clue in the sentence suggesting that the writer's attitude towards or judgment of the subject is both positive and negative,
- *Unknown* – there is no explicit or implicit clue indicating that the writer feels positively or negatively about the subject (this value is not used during the annotation process, whatever text that remains after annotating is considered unknown)

Each annotated sentence can only have 1 sentiment (Positive, Mixed, or Negative – these are disjoint). However, these sentences may also include a surprise emotion, such as when there is a sentence containing "I didn't know", or "it was nothing like I imagined", etc. In that case, they should also be annotated with the *surprised* emotion code. In this sentiment annotation phase, a total of 1878 annotations were created.

3.6 Sentiment analysis: employed techniques

To evaluate the quality of the annotations, we employed a technique called interannotator agreement. Interannotator agreement is a measure of how well the annotations of multiple annotators align with each other. This can be used to assess the consistency and reliability of the annotations. By measuring the interannotator agreement, we were able to determine whether the annotators were interpreting the task in the same way and to what extent their annotations matched.

Another technique employed in this experiment was the use of the Natural Language Toolkit (NLTK), a popular Python library for natural language processing. The NLTK library includes tools for sentiment analysis, which can be used to automatically classify text as positive, negative, or neutral. By using NLTK to generate sentiment scores for the annotated text, we were able to compare the sentiment scores to the annotations provided by the annotators and assess the accuracy of the annotations.

In addition to the techniques mentioned earlier, we also employed the use of TF-IDF (term frequency-inverse document frequency) in this experiment. TF-IDF is a statistical measure that is often used in information retrieval and natural language processing tasks to determine the importance of a word in a document. It is calculated by multiplying the term frequency (TF) of a word, which is the number of times the word appears in the document, by the inverse document frequency (IDF) of the word, which is the logarithm of the ratio of the total number of documents to the number of documents where the word appears.

In this experiment, we used TF-IDF to see if the words with the highest TF-IDF scores were indicative of sentiment or topic. The idea was that words with high TF-IDF scores would be the most important or relevant words in the document and therefore, may be more likely to carry sentiment or indicate a specific topic. Despite this, we still included TF-IDF as a technique in our experiment as it can be useful for other types of NLP tasks and provide valuable insight.

3.7 Sentiment analysis: results

Only a subset (half) of the stories were annotated by two annotators. This limited the amount of data available for evaluating interannotator agreement. Despite this limitation, we were able to calculate the interannotator agreement for the annotated stories.

The results revealed that the interannotator agreement among the annotators was not very high. This suggests that there was a lack of consistency in the way the annotators interpreted the task and made their annotations. This could be due to a variety of factors such as differences in the annotators' background, experience, or interpretation of the task instructions.

However, when the annotators did mark the same sentence, they often chose the same sentiment or aspect. This suggests that when the annotators were in agreement, they were likely interpreting the task in a similar way and making annotations that were consistent with each other.

Additionally, when comparing the sentiment scores generated by NLTK and the annotations provided by the annotators, we found that the sentiment scores generated by NLTK were not always in agreement with the annotations provided by the annotators. This highlights the limitations of using automated sentiment analysis tools and the importance of manual annotation to ensure a more accurate and nuanced analysis of the text.

Overall, the results of this experiment demonstrate the importance of interannotator agreement in evaluating the consistency and reliability of annotations. Despite the limitations of the data, the results suggest that when the annotators were in agreement, they were likely interpreting the task in a similar way and making annotations that were consistent with each other.

The following table represents the total number of sentiment annotations collected during the experiment.

	Positive	**Negative**	**Mixed**	**Surprised**
count	1060	545	218	55

Table 3 Sentiment annotations

The following table represents the confusion matrix of the two annotators, with the rows representing the annotation of annotator A and the columns representing the annotation of annotator B. Each cell of the table represents the number of times the annotators agreed on the annotation for a given sentiment (Positive, Negative, Mixed, Surprised).

	Positive	**Negative**	**Mixed**	**Surprised**
Positive	180	2	18	14
Negative	X	100	20	5
Mixed	X	X	23	1
Surprised	X	X	X	4

Table 4 Sentiment annotations confusion matrix

The following table shows the average sentiment score per word in sentences marked by both annotators as having the same sentiment, as scored by NLTK SentiWordNet.

	Positive	**Negative**	**Mixed**	**Surprised**
count	0.00523	- 0.00049	0.00376	0.00349

Table 5 Sentiment annotations interannotator agreement

When exploring the second, aspect-oriented set of annotations, it became clear that the annotators rarely chose the same segment of text. The following table represents the total number of basic aspects annotated.

	University life	**Private life**	**Life after PhD**	**Well-being**
count	439	97	54	165

Table 6 Aspect annotations

The following table represents the confusion matrix of the two annotators, with the rows representing the annotation of annotator A and the columns representing the annotation of

annotator B. Each cell of the table represents the number of times the annotators agreed on the annotation for a given aspect (University life, Private life, Life after PhD, Well-being).

	University life	**Private life**	**Life after PhD**	**Well-being**
University life	1	1	0	0
Private life	X	11	4	0
Life after PhD	X	X	30	0
Well-being	X	X	X	0

Table 7 Aspect annotations confusion matrix

4 Goal-directed analysis

4.1 Data sources

4.2 Modeling guidelines

4.3 Produced concept graphs

accessibility of Semantic Scholar in ContextMinds

5 Discussion

6 Conclusion

Future: fetching data from DBLP and similar databases

Acknowledgements

This research was supported by VSE IGS F4/56/2021. The authors are indebted to the volunteer annotators, without whom the research described in Section 3 would not have been possible.

References

Nguyen, V. B., V. Svátek, M. Dudás, and Ó. Corcho
2021. Knowledge engineering of PhD stories: A preliminary study. In *K-CAP '21: Knowledge Capture Conference, Virtual Event, USA, December 2-3, 2021*, A. L. Gentile and R. Gonçalves, eds., Pp. 281–284. ACM.