Improving Wikipedia Knowledge Exploration Capabilities by Similarity and Aspect Based Navigation

Master's Thesis Kick-Off Presentation

Emanuel Johannes Gerber, 23.5.2022

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

1. Motivation
2. Research Questions
3. Methods
   a. Semantic text similarity & aspect based text similarity
   b. User study
4. Timeline

Motivation

Goal Statement: Improving knowledge exploration by **automatically creating aspect specific links** between sections from Wikipedia

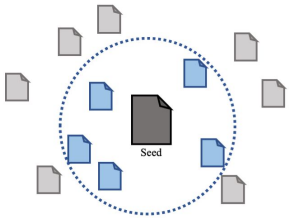
Knowledge Exploration

The process of obtaining insights within a new domain which is generally directed towards a complex, open-ended task.

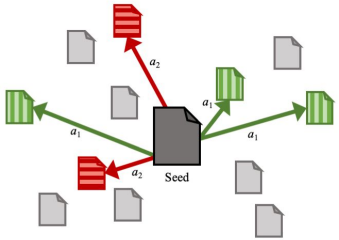
- E.g. “Find out about the Chinese Technology Sector”

Aspect Based Text Similarity

The similarity between texts that is specific towards a particular aspect/property.

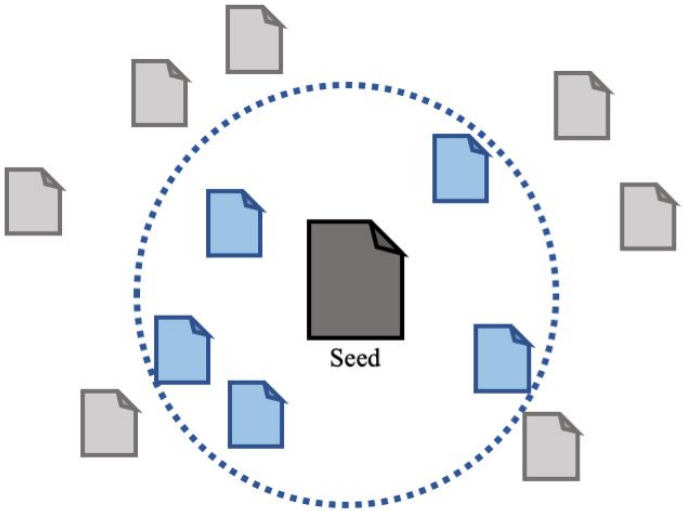


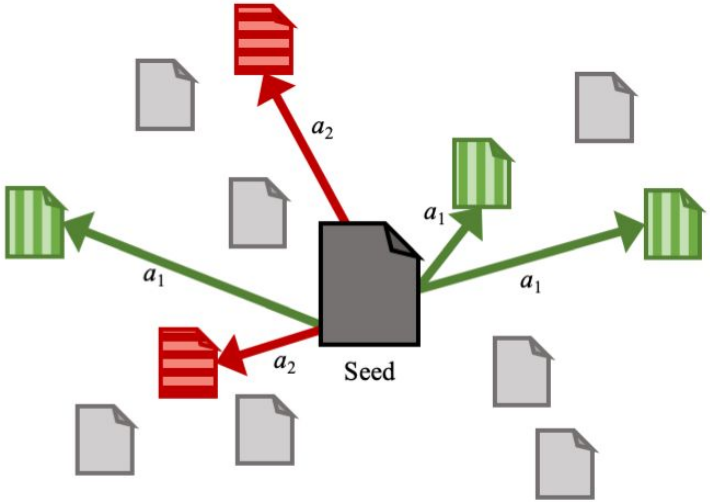
Generic Similarity



Aspect Based Similarity

Illustration: Aspect Based Similarity





Generic Similarity

Aspect Based Similarity

Research Questions

Q1

What is aspect based similarity?

Q2

What is the state of the art for semantic textual similarity and aspect-based similarity?

Q3

How to create embedding representations that capture the similarity for specific aspects?

Q4

How to evaluate the improvements in knowledge exploration on Wikipedia?

Methods: (Generic) Document Similarity

Word Embeddings

- Compute average of word embeddings avg_{di} for each document d_i (e.g. using *Word2vec or GloVe*)
- **similarity(d_1,d_2) = cosine_distance(avg_{d1}, avg_{d2})**

Key Phrases

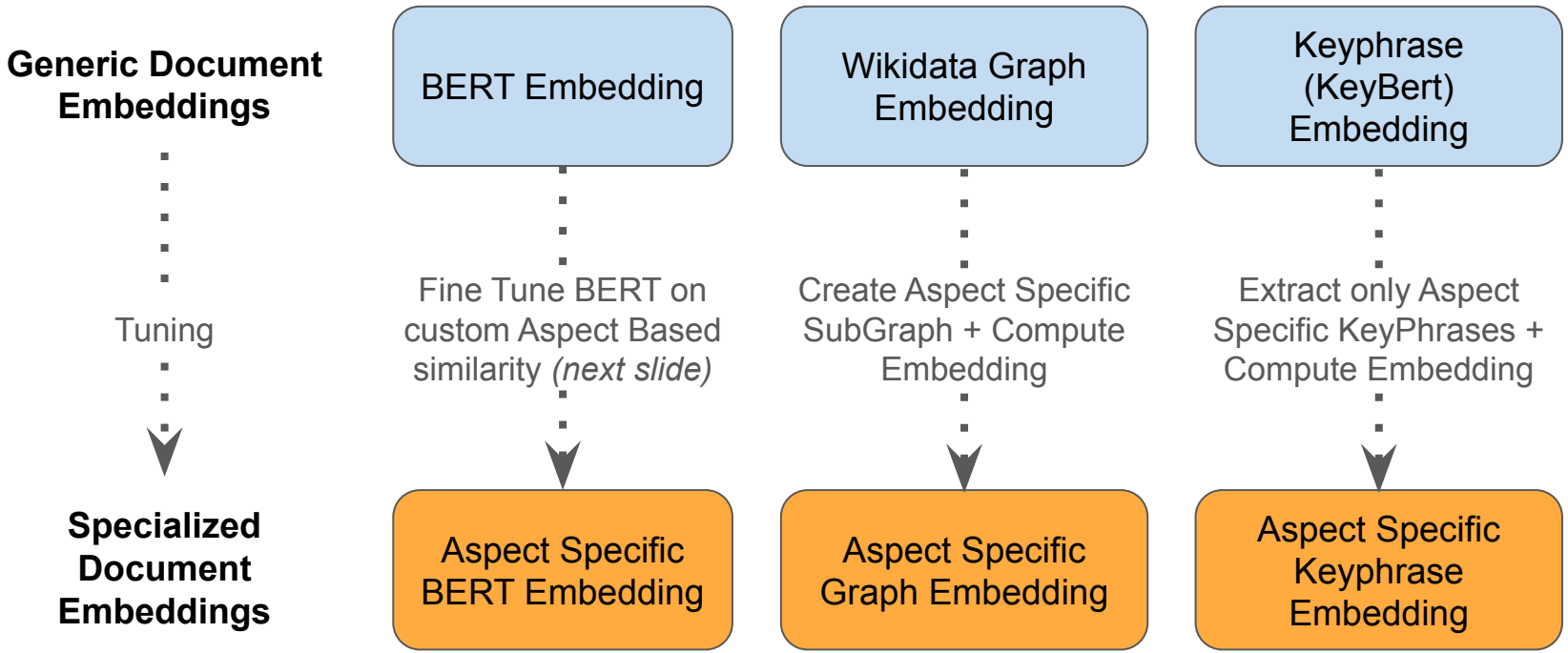
- Extract top k key phrases for each document d_i and create avg document embedding avg_d for all key phrases (e.g. using *KeyBert*)
- **similarity(d_1,d_2) = cosine_distance(avg_{d1}, avg_{d2})**

Document Embeddings

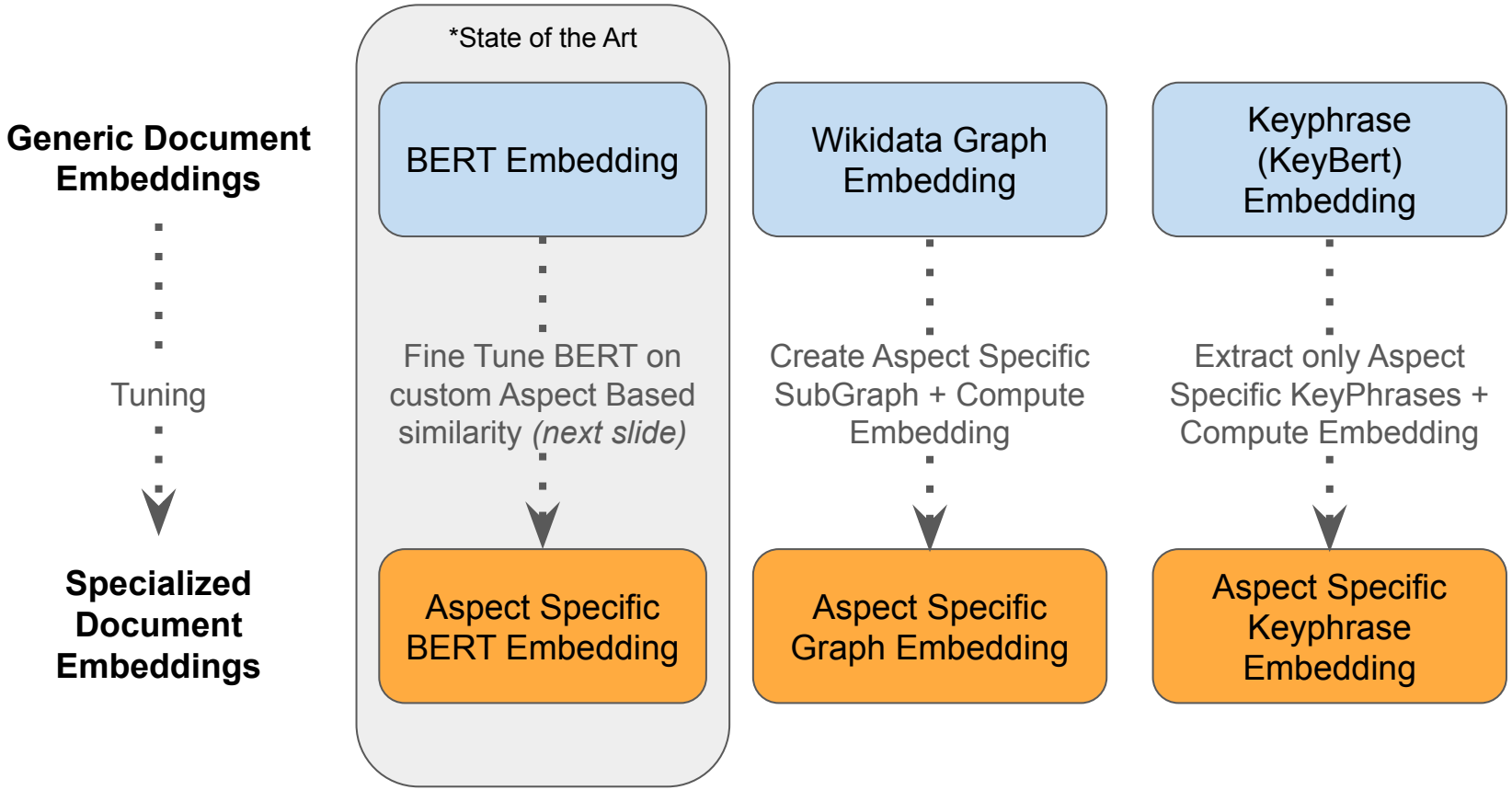
** State of the Art*

- Compute document embedding e_i for each document d_i using *SBERT*
- **similarity(d_1,d_2) = cosine_distance(e_1,e_2)**

Methods: (Aspect Based) Similarity



Methods: (Aspect Based) Similarity



Example: Creating Aspect Specific Embeddings

1. Define Aspects
   1.1. A1: “Technology related”
   1.2. A2: “China related”
   1.3. A3: “Industry related”
   1.4. A4: A1 ∩ A2 ∩ A3
2. Extract Aspect specific Documents/Sections from Wikipedia
   2.1. Using Text Matching (e.g. Text includes the word “China”) - naive approach
   2.2. Using WikiData Knowledge Graph (e.g. Entity is less than k links away from “China Entity”)
   2.3. Combinations of Keyword Extraction and Word Embeddings
3. Create Dataset of Triplets (d_1, d_2, y^a)
   3.1. y^a is {0,1} depending on if d_1 and d_2 share the same aspect a
4. Train Aspect Specialized Document Embedding (based on pretrained BERT embedding)
   4.1. Training Objective: maximize the similarity of the embeddings of document pairs (d_1, d_2) with y^a=1 (documents that are similar in aspect a)

** Derived from Ostendorf et. al 2022*

Methods: User Study

Goal Statement

"Find out about the Chinese Technology Sector" (complex, open ended task)

User Choices

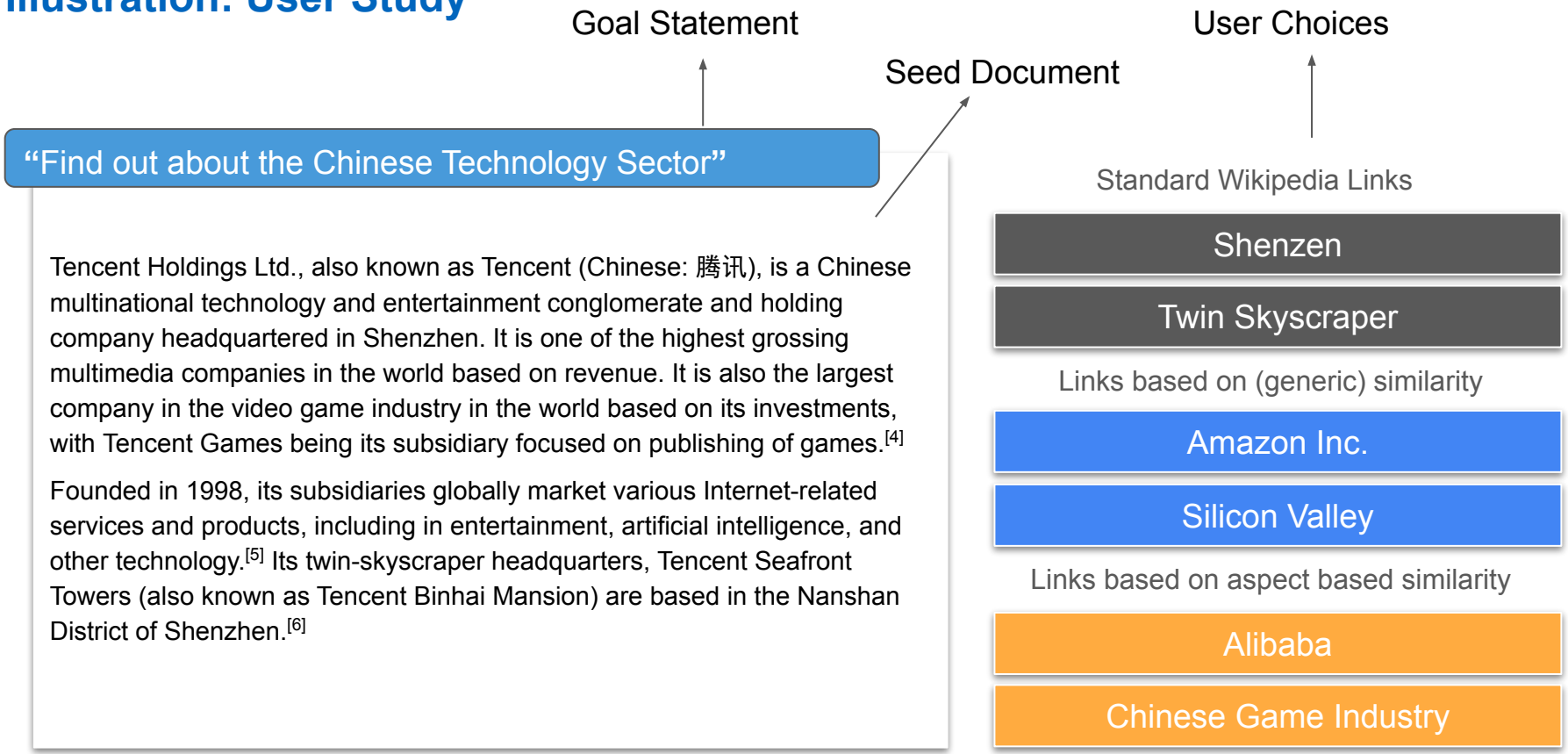
Given a seed article: Decide which text to visit next from the selection of…

- Wikipedia links
- Generic similarity based links
- Aspect based similarity links

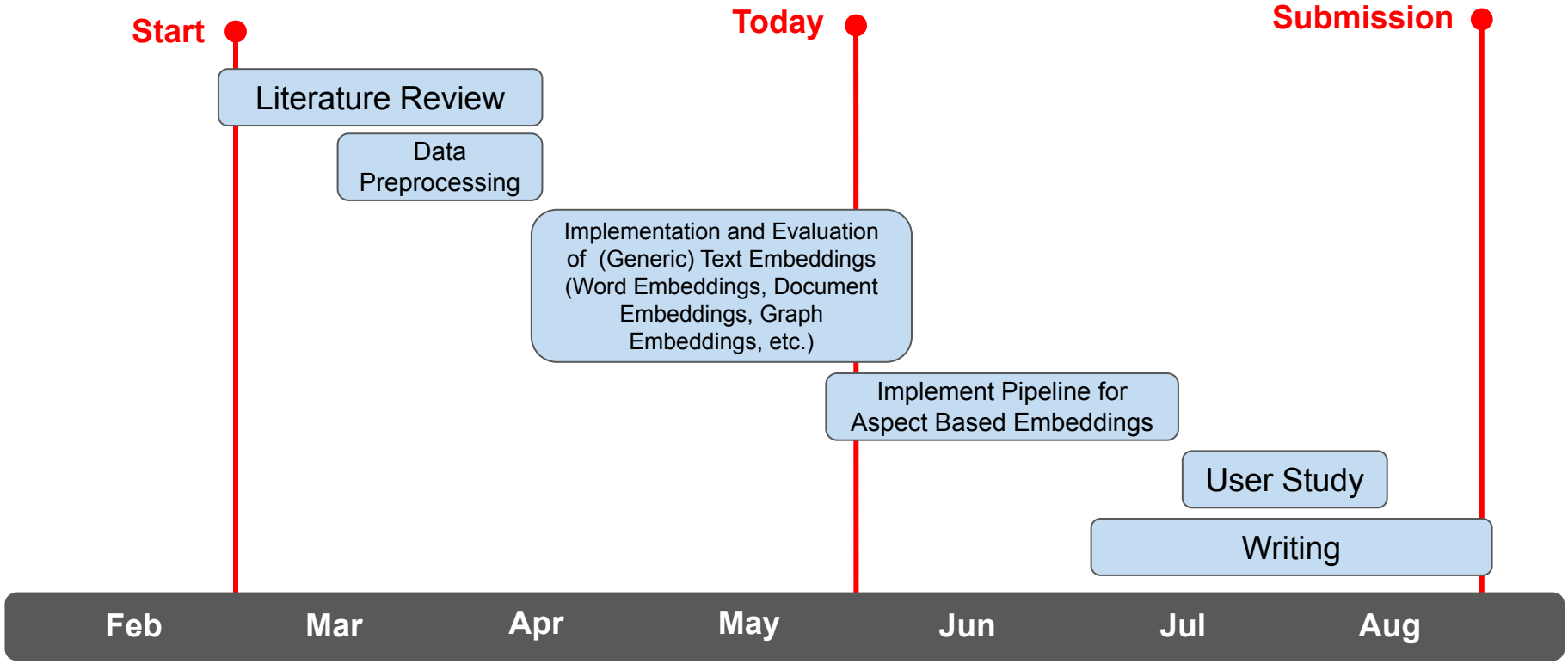
Metrics

- Frequency of choosing (aspect based) similarity link vs Wikipedia link
- "How helpful did you find the information that you received? (Scale 1-5)
- "How familiar are you with the information from this article?" (Scale 1-5)

Illustration: User Study



Timeline



Prof. Dr.
Florian Matthes

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for
Business Information Systems

Boltzmannstraße 3
85748 Garching bei München
17132

Tel +49.89.289.
Fax +49.89.289.17136

matthes@in.tum.de

www.matthes.in.tum.de