

### **Outline**

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

2

#### **Motivation**

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

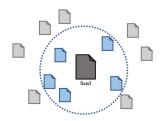
#### **Knowledge Exploration**

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

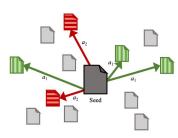
E.g. "Find out about the Chinese Technology Sector"

#### **Aspect Based Text Similarity**

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

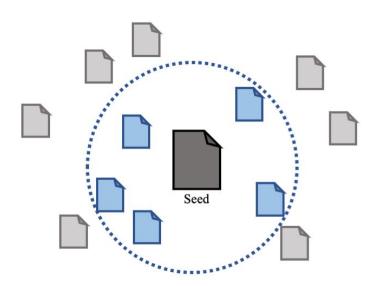


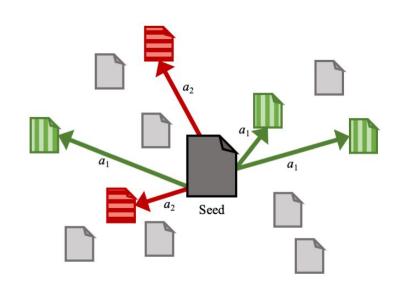
**Generic Similarity** 



**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<sub>di</sub> for each document d<sub>i</sub> (e.g. using *Word2vec or GloVe*)
- similarity(d<sub>1</sub>,d<sub>2</sub>) = cosine\_distance(avg<sub>d1</sub>, avg<sub>d2</sub>)

### **Key Phrases**

- Extract top k key phrases for each document d<sub>i</sub> and create avg document embedding avg<sub>d</sub> for all key phrases (e.g. using KeyBert)
- similarity(d<sub>1</sub>,d<sub>2</sub>) = cosine\_distance(avg<sub>d1</sub>, avg<sub>d2</sub>)

#### **Document Embeddings**

\* State of the Art

- Compute document embedding e, for each document d, using SBERT
- similarity( $d_1, d_2$ ) = cosine\_distance( $e_1, e_2$ )

## **Methods: (Aspect Based) Similarity**

### Generic Document Embeddings

:

**Tuning** 



Specialized Document Embeddings

### **BERT Embedding**

Fine Tune BERT on custom Aspect Based similarity (next slide)

Aspect Specific BERT Embedding

Wikidata Graph Embedding

Create Aspect Specific SubGraph + Compute Embedding

Aspect Specific Graph Embedding

Keyphrase (KeyBert) Embedding

Extract only Aspect Specific KeyPhrases + Compute Embedding

> Aspect Specific Keyphrase Embedding

## **Methods: (Aspect Based) Similarity**

### Generic Document Embeddings

:

Tuning



Specialized Document Embeddings

\*State of the Art

**BERT Embedding** 

Fine Tune BERT on custom Aspect Based similarity (next slide)

Aspect Specific BERT Embedding

Wikidata Graph Embedding

Create Aspect Specific SubGraph + Compute Embedding

Aspect Specific Graph Embedding

Keyphrase (KeyBert) Embedding

Extract only Aspect Specific KeyPhrases + Compute Embedding

> Aspect Specific Keyphrase Embedding

## **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<sub>1</sub>, d<sub>2</sub> y<sup>a</sup>)
  - 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)

<sup>\*</sup> 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**

**Goal Statement** 

**User Choices** 

**Seed Document** 

"Find out about the Chinese Technology Sector"

Tencent Holdings Ltd., also known as Tencent (Chinese: 腾讯), is a Chinese multinational technology and entertainment conglomerate and holding company headquartered in Shenzhen. It is one of the highest grossing multimedia companies in the world based on revenue. It is also the largest company in the video game industry in the world based on its investments, with Tencent Games being its subsidiary focused on publishing of games.<sup>[4]</sup>

Founded in 1998, its subsidiaries globally market various Internet-related services and products, including in entertainment, artificial intelligence, and other technology.<sup>[5]</sup> Its twin-skyscraper headquarters, Tencent Seafront Towers (also known as Tencent Binhai Mansion) are based in the Nanshan District of Shenzhen.<sup>[6]</sup>

Standard Wikipedia Links

Shenzen

Twin Skyscraper

Links based on (generic) similarity

Amazon Inc.

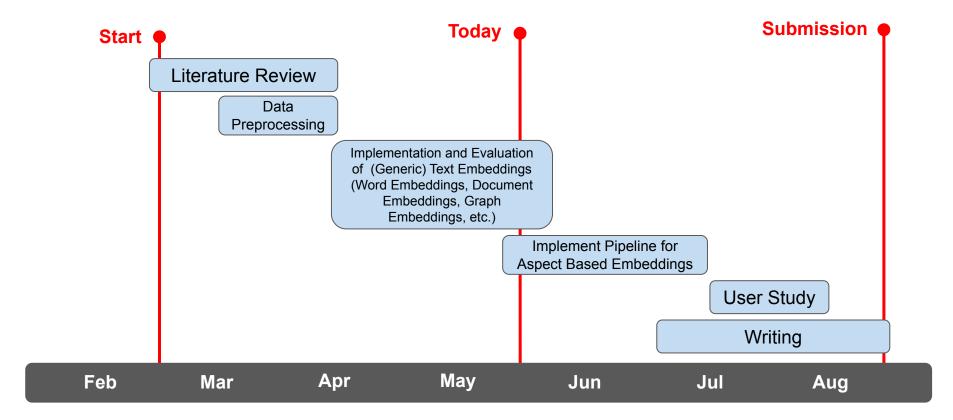
Silicon Valley

Links based on aspect based similarity

Alibaba

**Chinese Game Industry** 

### **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

