sebis

TUM

Improving Wikipedia Knowledge Exploration Capabilities by Similarity and Aspect Based Navigation

Master's Thesis Final Presentation

Emanuel Johannes Gerber, 13.09.2022

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. Experiments
4. Results

Motivation

Goal Statement

Improving **knowledge exploration** for Wikipedia articles by automatically creating **aspect specific links** between Wikipedia pages which address a particular information need of the user.

Knowledge Exploration

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

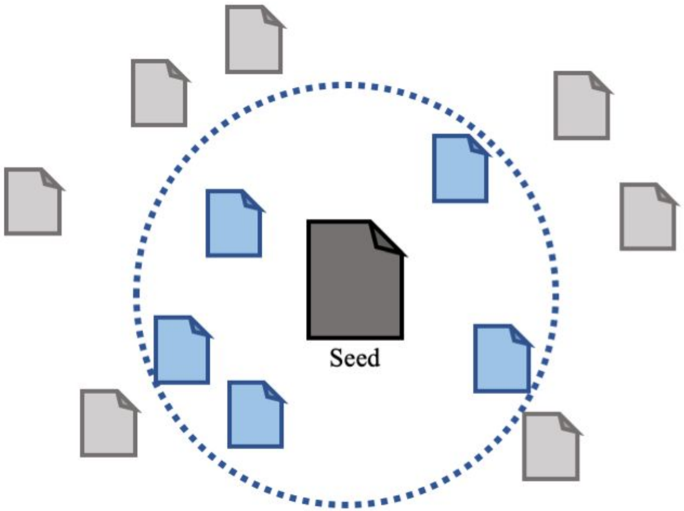
- E.g. “Find out about the Chinese Technology Sector”

Aspect Based Text Similarity

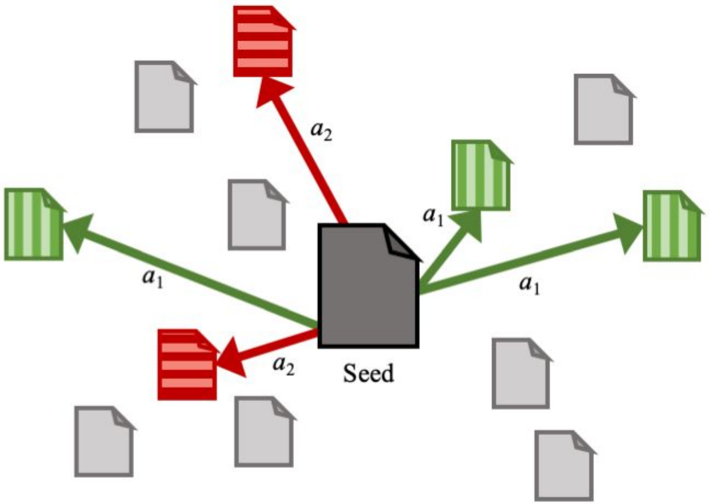
The similarity between texts that is specific towards a specific aspect

- E.g. two texts are similar if their subject is a Technology company.
- Or: two texts are similar if their subject is a Chinese company.

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?

Experiments

Use Cases

“Aspect-Based Document Embeddings have not been applied to Wikipedia so far”

Papers With Code

PapersWithCode

*NEW

Wikipedia

*NEW

Data Augmentation

Training Data Composition

“Aspect-Based Document Embeddings require a lot of training data”

WIKIDATA

*NEW

External Knowledge Bases

Models

“New State-of-the-Art Methods for Training (Generic) Embeddings have been developed”

SBERT.net

SBERT

princeton-nlp/
SimCSE

*NEW

Contrastive Learning
(SimCSE)

Data Augmentation

*"**Data augmentation** in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data."* (Wikipedia)

Original Text

room classification on floor plan graphs using graph neural networks. we present our approach to improving room classification tasks on floor plan maps of buildings by representing floor plans as undirected graphs and leveraging graph neural networks to predict the room categories. rooms in the floor plans are represented as nodes in the graph with edges representing their adjacency in the map. we experiment with house-gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks. our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80% and 81% respectively outperforming baseline multilayer perceptron by more than 15% margin.

Augmented Text

Summarization

room classification task on floor plan maps of buildings using graph neural networks. rooms in the floor plans are represented as nodes in a graph with edges indicating their adjacency in map - our approach aims to improve room category prediction based on graphsage and topology adaptive gcn models

Data Augmentation

*"**Data augmentation** in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data."* (Wikipedia)

Original Text

room classification on floor plan graphs using graph neural networks. we present our approach to improving room classification tasks on floor plan maps of buildings by representing floor plans as undirected graphs and leveraging graph neural networks to predict the room categories. rooms in the floor plans are represented as nodes in the graph with edges representing their adjacency in the map. we experiment with house-gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks. our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80% and 81% respectively outperforming baseline multilayer perceptron by more than 15% margin.

Augmented Text

Synonym Replacement

room classification on floor plan graphs using graph optical networks. we present our technique to improve room classification task using floor plan maps of houses by representing floor plans as undirected graphs and lever-aging graph optic networks to infer the room categories. rooms in the floor plans are represented using nodes in the graph with edges representing their neighborhood in the map. we experiment with house - gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks. our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80 % and 81 % respectively outperforming baseline multilayer perceptron by more than 15 % margin

Data Augmentation

*“**Data augmentation** in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data.”* (Wikipedia)

Original Text

room classification on floor plan graphs using graph neural networks. we present our approach to improving room classification tasks on floor plan maps of buildings by representing floor plans as undirected graphs and leveraging graph neural networks to predict the room categories. rooms in the floor plans are represented as nodes in the graph with edges representing their adjacency in the map. we experiment with house-gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks. our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80% and 81% respectively outperforming baseline multilayer perceptron by more than 15% margin.

Augmented Text

Sentence Shuffling

we experiment with house-gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks.rooms in the floor plans are represented as nodes in the graph with edges representing their adjacency in the map.room classification on floor plan graphs using graph neural networks.our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80% and 81% respectively outperforming baseline multilayer perceptron by more than 15% margin. we present our approach to improve room classification task on floor plan maps of buildings by representing floor plans as undirected graphs and leveraging graph neural networks to predict the room categories.

Data Augmentation

*"**Data augmentation** in data analysis are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data."* (Wikipedia)

Original Text

room classification on floor plan graphs using graph neural networks. we present our approach to improving room classification tasks on floor plan maps of buildings by representing floor plans as undirected graphs and leveraging graph neural networks to predict the room categories. rooms in the floor plans are represented as nodes in the graph with edges representing their adjacency in the map. we experiment with house-gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks. our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80% and 81% respectively outperforming baseline multilayer perceptron by more than 15% margin.

Augmented Text

Sentence Deletion

room classification on floor plan graphs using graph neural networks. we experiment with house-gan dataset that consists of floor plan maps in vector format and train multilayer perceptron and graph neural networks. our results show that graph neural networks, specifically graphsage and topology adaptive gcn were able to achieve accuracy of 80% and 81% respectively outperforming baseline multilayer perceptron by more than 15% margin.

PapersWithCode - Data Augmentation Results

Improvements by avg. **~4%** for MRR

Improvements through Data Augmentation by **~6%** for MRR

Existing Benchmark by Ostendorff et. al 2022

Generic Text Embeddings

Aspects →	*Task*			*Method*			*Dataset*		
Methods ↓	**P**	**R**	**MRR**	**P**	**R**	**MRR**	**P**	**R**	**MRR**
SimCSE									
Gold	0.416	0.431	0.776	0.268	0.312	0.606	0.186	0.461	0.507
CWE Substitution	0.400	0.417	0.759	0.262	0.303	0.586	0.186	0.455	0.515
Summarization	0.396	0.416	0.742	0.260	0.297	0.585	0.188	0.463	0.541
Sentence Deletion	0.394	0.410	0.711	0.261	0.291	0.581	0.188	0.466	0.528
SBERT									
Gold*	0.409	0.424	0.768	0.263	0.302	0.595	0.172	0.418	0.465
CWE Substitution	0.407	0.419	0.751	0.262	0.301	0.592	0.170	0.411	0.461
Summarization	0.396	0.416	0.742	0.244	0.285	0.566	0.171	0.416	0.463
Sentence Deletion	0.395	0.411	0.736	0.244	0.284	0.556	0.168	0.412	0.458
Lower Bound	0.089	0.152	0.048	0.400	0.039	0.124	0.119	0.316	0.072

Training Data Composition with Wikidata

WIKIDATA

Tencent

instance of	holding company (0 references)
	conglomerate (0 references)
	public company (0 references)
industry	video game industry (0 references)
	information technology (0 references)
country	People's Republic of China (1 reference)

Alibaba

instance of	holding company (0 references)
	conglomerate (0 references)
	public company (0 references)
industry	e-commerce (1 reference)
country	People's Republic of China (1 reference)

Electronic Arts

instance of	video game publisher (1 reference)
	public company (0 references)
industry	video game industry (1 reference)
country	United States of America (1 reference)

Training Data Composition with Wikidata

WIKIDATA

Country Aspect

Seed Document	Positive Document	Negative Document
Tencent	Alibaba	Electronic Arts

Industry Aspect

Seed Document	Positive Document	Negative Document
Tencent	Electronic Arts	Alibaba

Mixed Aspect (Industry OR Country)

Seed Document	Positive Document	Negative Document
Tencent	Alibaba	GAZPROM

Seed Document	Positive Document	Negative Document
Tencent	Electronic Arts	GAZPROM

Data Composition with Wikidata

Aspects →	*Country*			*Industry*		
Methods ↓	**P**	**R**	**MRR**	**P**	**R**	**MRR**
Single Aspect	0.390	0.124	0.558	0.625	0.178	0.729
Multi Aspect (Join)	0.444	0.102	0.593	0.622	0.174	0.720
Multi Aspect (Union)	0.555	0.163	0.738	0.538	0.155	0.747
Lower Bound	0.315	0.058	0.523	0.320	0.061	0.531

Outperforming Single Aspect Embeddings

Generic Text Embeddings

Illustration Aspect Based Document Embeddings

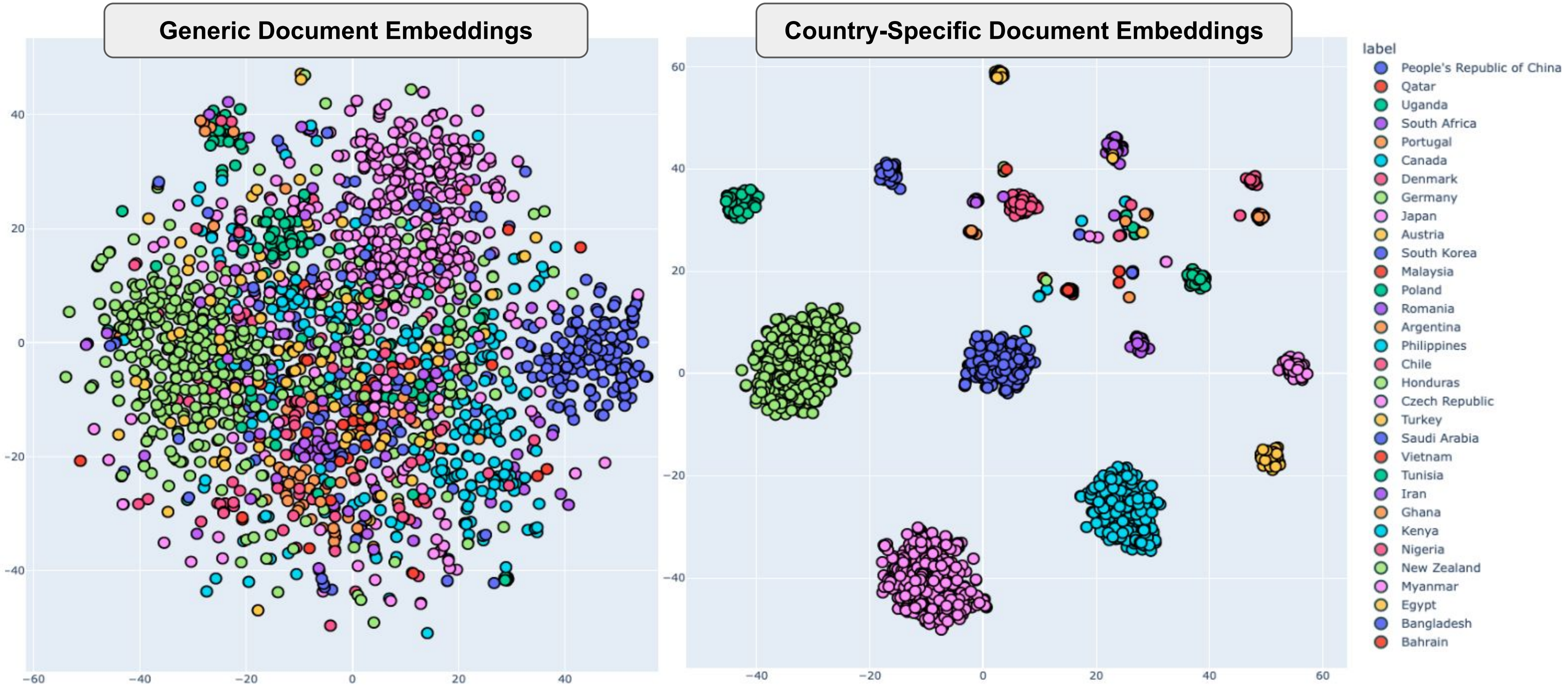
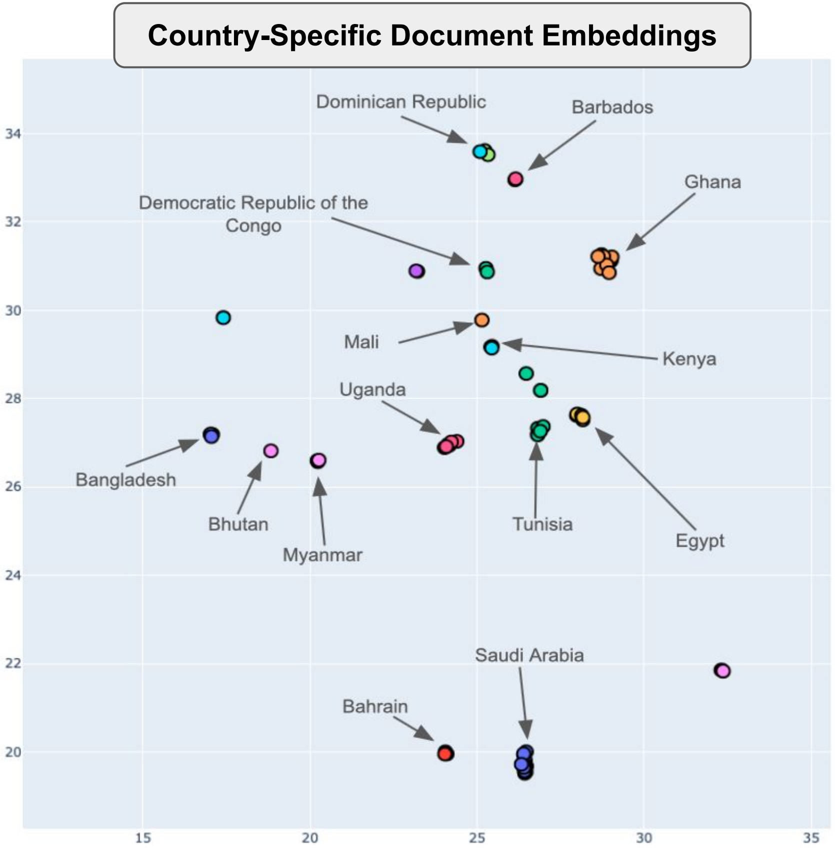


Illustration Aspect Based Document Embeddings



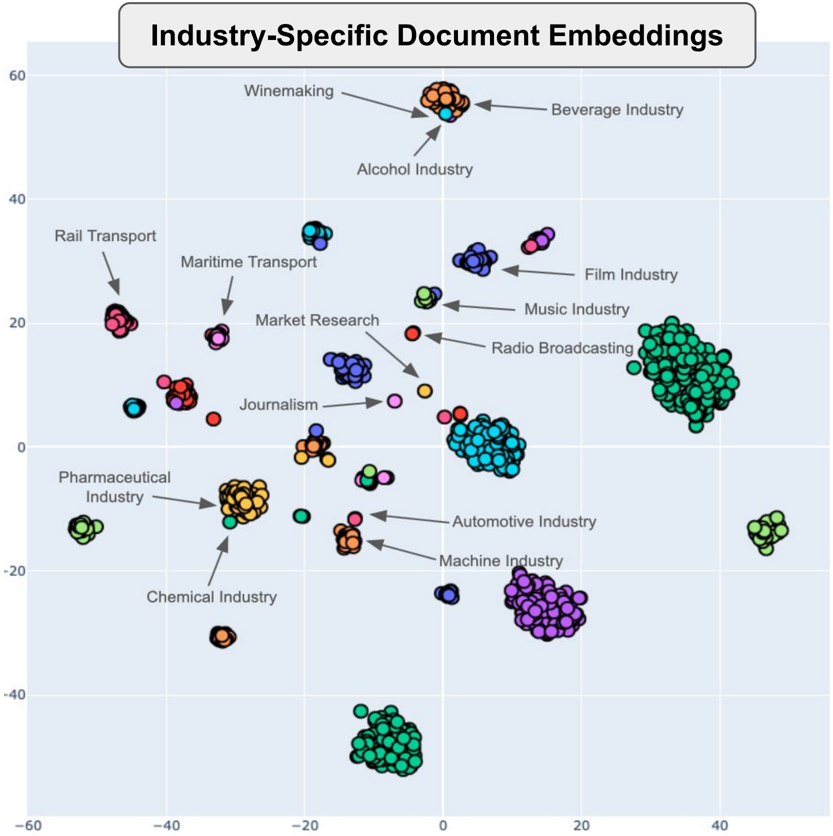
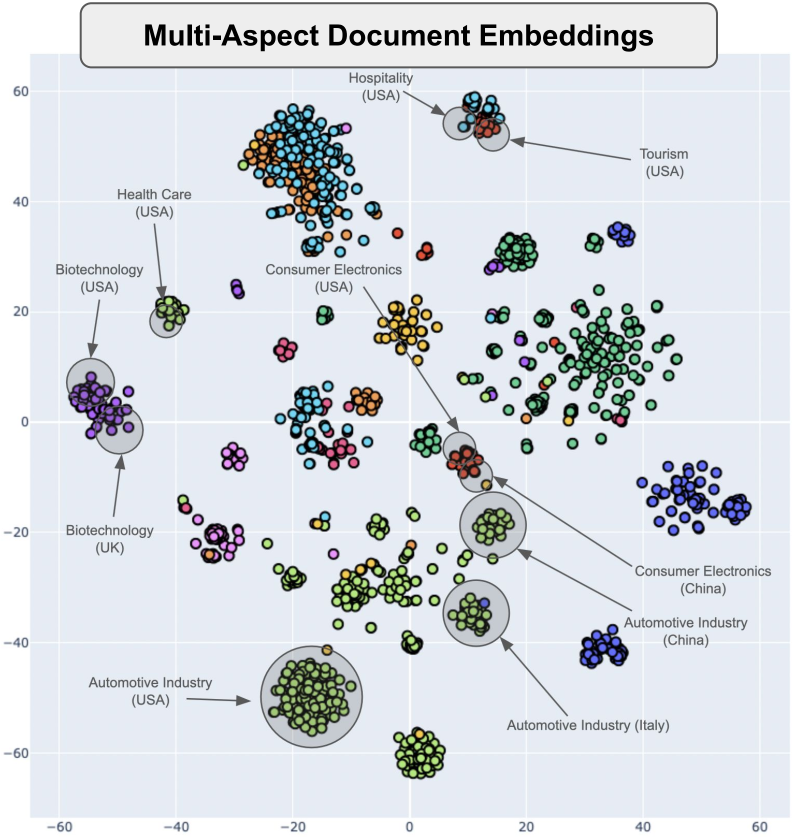


Illustration Aspect Based Document Embeddings



User Study

21 users

10 tasks

8-12 user choices per task

Randomization

=> ~2100 data points

User Prompt

Assume you want to learn more about E-Commerce Companies. Please read the given article and then rank which of the given articles you would like to read next (1 to 5)

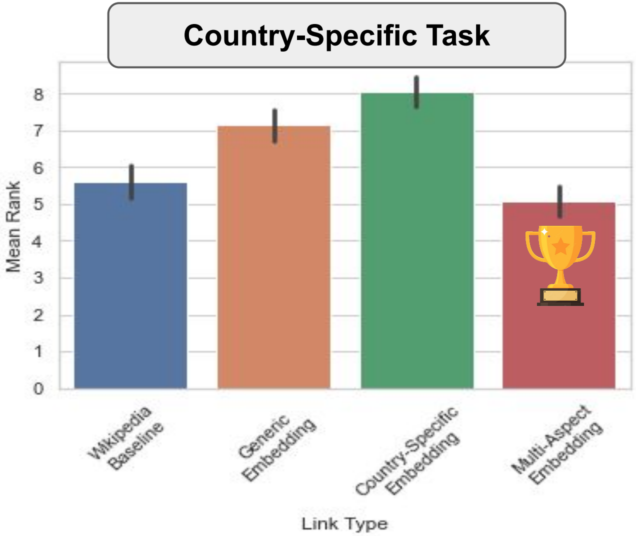
Amazon.com, Inc. is an American multinational technology company based in Seattle that focuses on e-commerce, cloud computing, digital streaming, and artificial intelligence. It is considered one of the Big Four tech companies, along with Google, Apple, and Facebook….

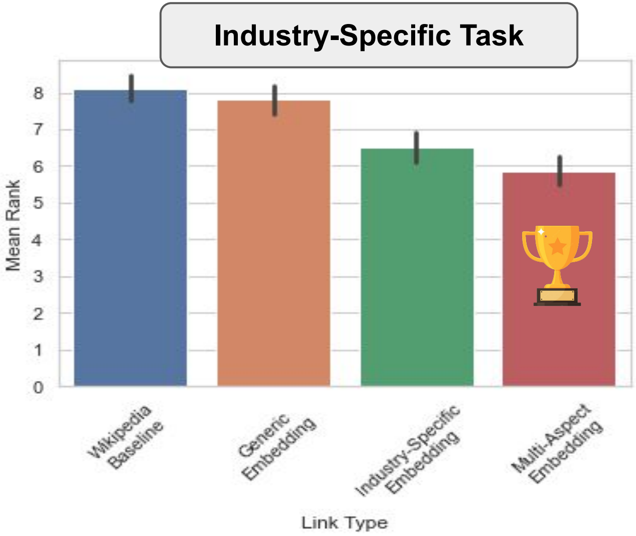
User Choices

User Choices	
Amazon Prime is a paid subscription service offered by Amazon that gives users access to services that would otherwise be unavailable, or cost extra, to the typical Amazon customer…	**Generic Text Embeddings**
TigerDirect is an El Segundo, California-based online retailer dealing in electronics, computers, and computer components that caters to business and corporate customers…	**Industry Specific Embedding**
Art Technology Group (ATG) was an independent Internet technology company specializing in eCommerce software and on-demand optimization applications until its acquisition by Oracle on January 5, 2011	**Multi-Aspect Embedding (Country + Industry)**
A technology company (or tech company) is an electronics-based technological company, including, for example, business relating to digital electronics, software, and internet-related services, such as e-commerce services	**Baseline Wikipedia Link**

…

User Study Results





Conclusion: Multi-Aspect Article Recommendations reflect a “more natural” user need

Conclusions

- Data Augmentation only works when **sufficiently enough** ground-truth data is given
- **Contrastive Learning** (SimCSE) **consistently outperforms SBERT** on aspect based embeddings
- **Wikidata is a rich source** for composing aspect based documents
- Multi-Aspect embeddings show **unexpected qualities**
- Users show a preference towards Multi-Aspect embeddings

Sentence Embeddings for Aspect-based Semantic Textual Similarity using Contrastive Learning and Structured Knowledge

Anonymous ACL submission

Abstract

Generic sentence embeddings only provide a coarse-grained approximation of semantic textual similarity and ignore the specific aspects that make texts similar. In contrast, aspect-based sentence embeddings provide similarities between texts with respect to certain predefined aspects. This allows similarity predictions of texts to be more targeted to specific requirements as well as to be more easily explainable. In this work, we show that using contrastive learning to train aspect-based sentence embeddings yields an average improvement of 3.3% on information retrieval tasks across multiple aspects compared to previous best results. In addition, we propose to use Wikidata Knowledge Graph relations to train multi-aspect sentence embedding models that consider multiple specific aspects simultaneously during similarity predictions. We demonstrate that multi-aspect embeddings outperform even single-aspect embeddings on aspect-specific information retrieval tasks. Finally, we examine the aspect-based sentence embedding space and show that embeddings of semantically similar aspect labels are often close, even without explicit training for similarity between different aspect labels.

1 Introduction

(Ostendorff et al., 2020a). Moreover, they are usually evaluated on generic Semantic Textual Similarity (STS) tasks (Marelli et al., 2014; Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017), which rely on human annotations of sentence similarity scores. However, the concept of generic STS is not well defined, and the similarity of texts depends heavily on the aspects making them similar (Bär et al., 2011; Ostendorff et al., 2020b, 2022). We follow the argumentation of Bär et al. (2011) on textual similarity and define *aspects* as inherent properties of texts that need to be considered when predicting their semantic similarity. Based on the different aspects focused on in texts, their similarities can be perceived very differently. Figure 1 illustrates an example of aspect-based STS. Looking, for example, at Wikipedia introduction texts of famous individuals, they can generally be considered similar, as all texts are about people who are known to the general public. However, focusing the comparison on specific aspects, such as *country of birth* or *profession*, leads to different semantic similarity assessments for the same texts. When deciding on how similar texts are, many different aspects can be considered. Consequently, human annotated STS datasets introduce a considerable amount of subjectivity with respect to the aspects being evaluated.

Next Steps

- Identify user needs in order to build custom user exploration
- Explore Sequence-To-Sequence models for composing custom aspect-based datasets
  (*Arxiv: Towards Zero-Label Language Learning)*

Q&A

Backup - Data Augmentation (100 instances per Label)

Aspects →	*Task*			*Method*			*Dataset*		
Methods ↓	**P**	**R**	**MRR**	**P**	**R**	**MRR**	**P**	**R**	**MRR**
Upper Bound	0.409	0.424	0.768	0.263	0.302	0.595	0.172	0.418	0.465
SimCSE									
Gold	0.290	0.332	0.626	0.196	0.238	0.488	0.173	0.438	0.509
backtranslation	0.288	0.332	0.625	0.172	0.214	0.433	0.164	0.427	0.514
CWE Insertion	0.295	0.334	0.631	0.178	0.218	0.449	0.169	0.432	0.509
CWE Substitution	0.295	0.333	0.631	<u>0.201</u>	0.241	<u>0.501</u>	<u>0.174</u>	<u>0.447</u>	<u>0.522</u>
PLS	0.246	0.291	0.557	0.127	0.158	0.317	0.173	0.438	0.516
Sentence Deletion	0.284	0.327	0.613	0.170	0.211	0.443	0.171	0.446	0.513
Sentence Reordering	<u>0.300</u>	<u>0.341</u>	<u>0.641</u>	0.200	<u>0.242</u>	0.500	0.170	0.441	0.510
Summarization	0.278	0.322	0.604	0.172	0.216	0.435	0.169	0.440	0.513
SBERT									
Gold	0.221	0.253	0.515	0.095	0.115	0.266	0.138	0.364	0.420
backtranslation	0.198	0.222	0.520	0.142	0.180	0.345	0.130	0.338	0.414
CWE Insertion	0.234	0.275	0.541	0.105	0.130	0.302	0.141	0.365	0.422
CWE Substitution	<u>0.235</u>	<u>0.270</u>	<u>0.543</u>	0.130	0.181	<u>0.381</u>	<u>0.140</u>	<u>0.365</u>	<u>0.425</u>
PLS	0.177	0.214	0.451	0.069	0.088	0.209	0.126	0.348	0.420
Sentence Deletion	0.215	0.241	0.508	0.117	0.146	0.323	0.119	0.341	0.402
Sentence Reordering	0.222	0.255	0.528	0.128	0.166	0.370	0.129	0.345	0.424
Summarization	0.230	0.269	0.536	<u>0.162</u>	<u>0.200</u>	0.335	0.136	0.360	<u>0.425</u>
Lower Bound	0.089	0.152	0.048	0.400	0.039	0.124	0.119	0.316	0.072

User Study - Example

soSci
oFb - der onlineFragebogen

59% completed

2. E-Commerce Companies

Assume that you want to learn more about **E-Commerce Companies**. Please read the given text and then rank the Top 5 articles that you would like to read next (1 indicates your top choice).

Alibaba Group Holding Limited, (also known as Alibaba Group and as Alibaba), is a Chinese multinational conglomerate holding company specializing in e-commerce, retail, Internet, and technology. Founded on 4 April 1999 in Hangzhou, Zhejiang, the company provides consumer-to-consumer (C2C), business-to-consumer (B2C), and business-to-business (B2B) sales services via web portals, as well as electronic payment services, shopping search engines and cloud computing services. It owns and operates a diverse array of businesses around the world in numerous sectors, and is named as one of the world's most admired companies by Fortune. At closing time on the date of its initial public offering (IPO) – US$25 billion – the world's highest in history, 19 September 2014, Alibaba's market value was US$231 billion.

CHANNELADVISOR: ChannelAdvisor Corp. is an e-commerce company based in Morrisville, North Carolina. The company provides cloud-based e-commerce software solutions to ... –

MULTINATIONAL CORPORATION: A multinational corporation or worldwide enterprise is a corporate organization that owns or controls production of goods or services in at leas... –

VIPSHOP: Vipshop is a Chinese company that operates the e-commerce website VIP.com specializing in online discount sales. Vipshop is based out of Guangzhou, Gu... –

ALIEXPRESS: AliExpress is an online retail service based in China that is owned by the Alibaba Group. Launched in 2010, it is made up of small businesses in China... –

KABAM: Kabam is an interactive entertainment company founded in 2006 and headquartered in Vancouver, BC. with offices in San Francisco, CA and Austin, Texas.... –

CONGLOMERATE (COMPANY): A conglomerate is a combination of multiple business entities operating in entirely different industries under one corporate group, usually involving ... –

TWENGA: Twenga co-founders Bastien Duclaux and Cédric Anès met in 2000 at the École Nationale Supérieure des Télécommunications. The company's "crawl" technol... –

TMALL: Tmall.com , formerly Taobao Mall, is a Chinese-language website for business-to-consumer online retail, spun off from Taobao, operated in China by Ali... –

HOLDING COMPANY: A holding company is a company that owns other companies' outstanding stock. A holding company usually does not produce goods or services itself (with... –

JD.COM: JD.com, Inc. , also known as Jingdong and formerly called 360buy, is a Chinese e-commerce company headquartered in Beijing. It is one of the two massi... –

TAOBAO: Taobao is a Chinese online shopping website, headquartered in Hangzhou, and owned by Alibaba. It is the world's biggest e-commerce website and the eig... –

EBAY: eBay Inc. is an American multinational e-commerce corporation based in San Jose, California, that facilitates consumer-to-consumer and business-to-con... –

Backup - 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)**

Backup - 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*

Illustration: User Study

Goal Statement

Seed Document

User Choices

“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.[4]

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

Standard Wikipedia Links

- Shenzhen
- Twin Skyscraper

Links based on (generic) similarity

- Amazon Inc.
- Silicon Valley

Links based on aspect based similarity

- Alibaba
- Chinese Game Industry

Datasets

Papers With Code

- Experimenting with Data Augmentation
- Comparing Model Performance with existing Benchmark

Wikipedia
The Free Encyclopedia

- Constructing Aspect-Based Datasets from External Knowledge Bases
- Multi-Aspect Embeddings