

Cross-Dialect Information Retrieval: Information Access in Low-Resource and High-Variance Languages

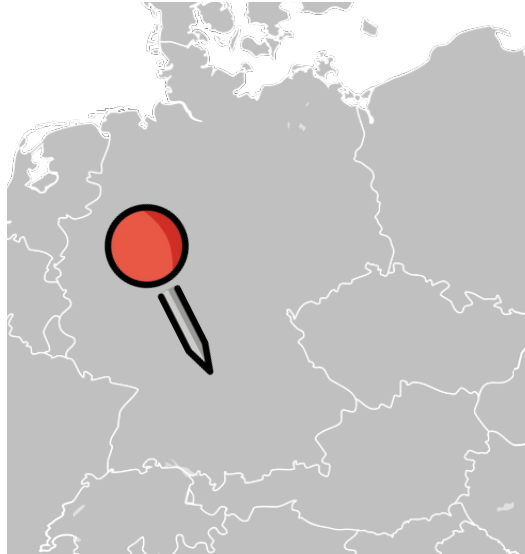
Robert Litschko, Oliver Kraus, Verena Blaschke, Barbara Plank



München (“Munich”)



München ("Munich")



Wikipedia
<https://de.wikipedia.org/wiki/München>

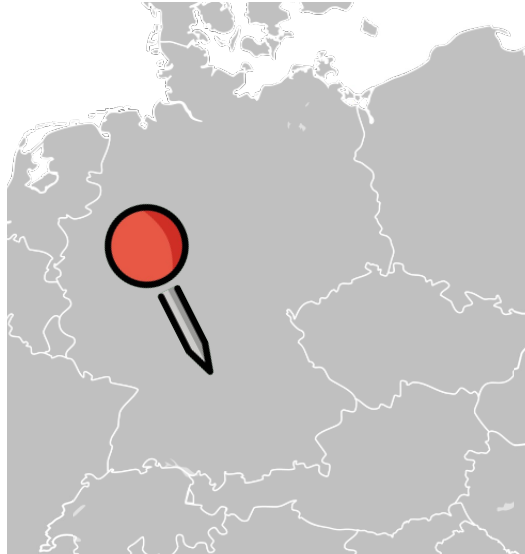
München

Sie ist mit gut 1,5 Millionen Einwohnern die bevölkerungsreichste
Gemeinde Deutschlands und mit 4.861 Einwohnern die kleinste
[Geschichte Münchens – Altstadt \(München\) – Land...](#)



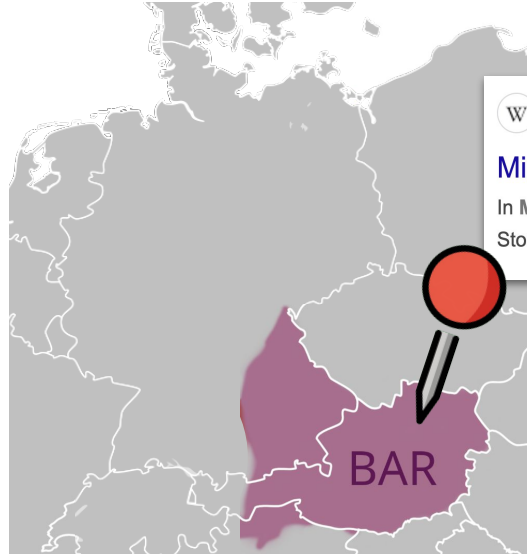
München ("Munich")



What about **culture-specific knowledge** that can often be found in dialect Wikis?





München ("Munich")



 Boarische Wikipedia
<https://bar.wikipedia.org/wiki/> · [Translate this page](#) · 

Minga

In Minga sogt ma München. Minga is mid mehra wia 1 Stod vo Bayern und hinta Berlin und Hamburg d'drittgre




München ("Munich")

 Alemannische Wikipedia
<https://als.wikipedia.org> · München · [Translate this page](#) · [More](#)

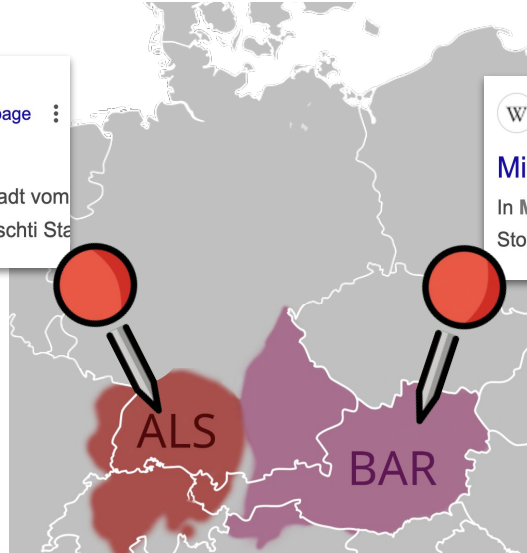
Münche

Münche (hd. **München**, bar. Minga) isch d Hauptstadt vom Bayern un mit über 1,4 Millione liwohner au die gröschti Sta

 Boarische Wikipedia
<https://bar.wikipedia.org> · wiki · [Translate this page](#) · [More](#)

Minga

In Minga sogt ma München. Minga is mid mehra wia 1 Stod vo Bayern und hinta Berlin und Hamburg d'drittgre





München ("Munich")



Alemannische Wikipedia

<https://als.wikipedia.org> · München · [Translate this page](#)

Münche

Münche (hd. **München**, bar. Minga) isch d Hauptstadt vom Bayern un mit über 1,4 Millione liwohner au die gröschti Sta

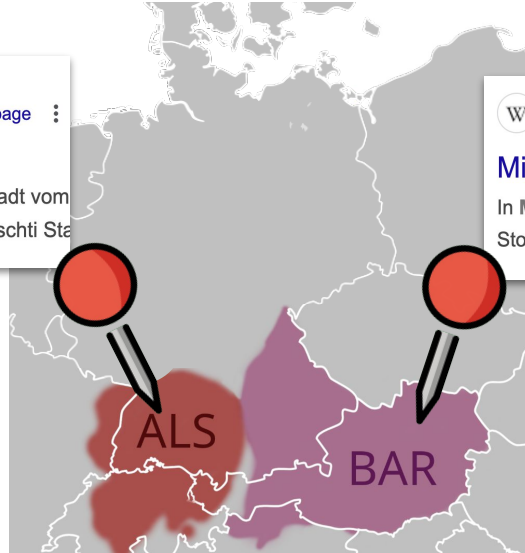


Boarische Wikipedia

<https://bar.wikipedia.org> · wiki · [Translate this page](#)

Minga

In Minga sogt ma München. Minga is mid mehra wia 1 Stod vo Bayern und hinta Berlin und Hamburg d'drittgre



Mincke Minche Mincha
Müncha Münchu
Minke Münchè
Minchen Münchä
Minga Mìncha Minchä

Minchn Minkhn
Münc'h'n
Minkcha Münchn
Minkn Mingna



High lexical variation due to regional word choices and different pronunciations.



Alemannische Wikipedia

<https://als.wikipedia.org> · München · [Translate this page](#)

Münche

Münche (hd. **München**, bar. Minga) isch d Hauptstadt vom Bayern un mit über 1,4 Millione liwohner au die gröschti Sta

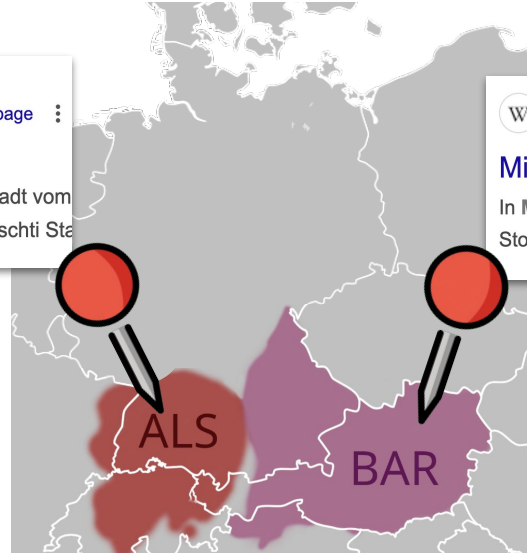


Boarische Wikipedia

<https://bar.wikipedia.org> · wiki · [Translate this page](#)

Minga

In Minga sogt ma München. Minga is mid mehra wia 1 Stod vo Bayern und hinta Berlin und Hamburg d'drittgre



Mincke Minche Mincha Münchu Münschen Münchèn Münchä Minke Minchen Minka Mìncha Minchä

Minchn Minkhn Münch'n Minkcha Minkn Mingna Münchn



High lexical variation due to regional word choices and different pronunciations.



Alemannische Wikipedia

<https://als.wikipedia.org> · München · [Translate this page](#)

Münche

Münche (hd. **München**, bar. Minga) isch d Hauptstadt vom Bayern un mit über 1,4 Millione liwohner au die gröschti Sta

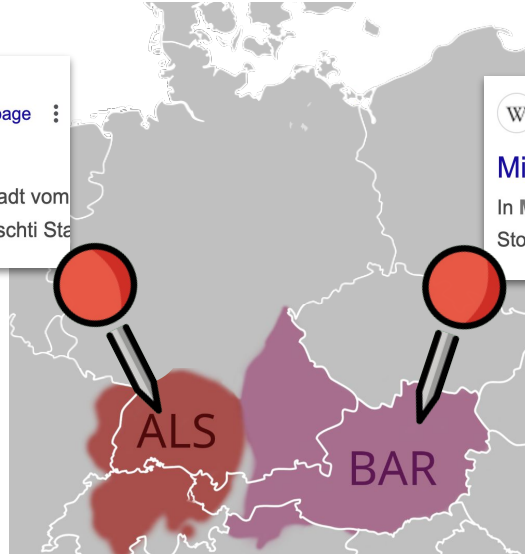


Boarische Wikipedia

<https://bar.wikipedia.org> · wiki · [Translate this page](#)

Minga

In Minga sogt ma München. Minga is mid mehra wia 1 Stod vo Bayern und hinta Berlin und Hamburg d'drittgre



Mincke Minche Mincha
Müncha Münchu
Minke Münchèn
Minchen Münchä
Minga Mìncha Minchä

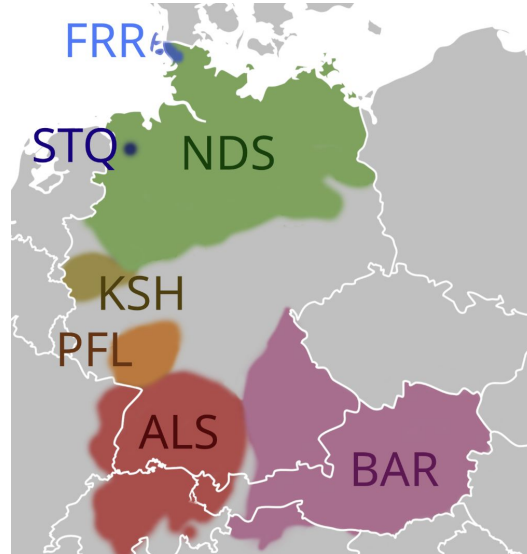
Minchn Minkhn
Münc'h'n
Minkcha Münchn
Minkn Mingna



Lexical retrieval falls short: Normalizers do not exists for most dialects.



High lexical variation due to regional word choices and different pronunciations.



Low German (**nds**)

Alemannic (**als**)

Bavarian (**bar**)

North Frisian (**frr**)

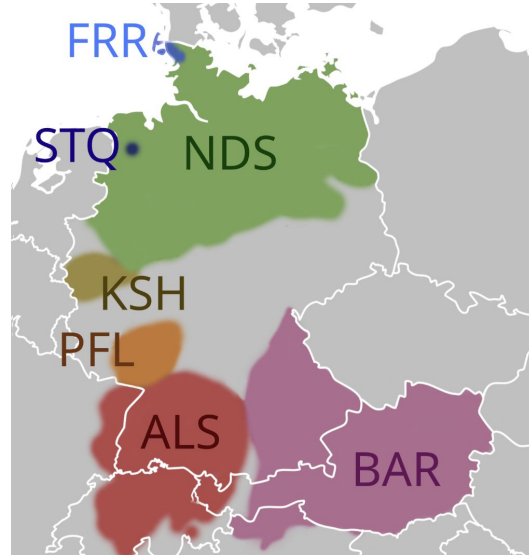
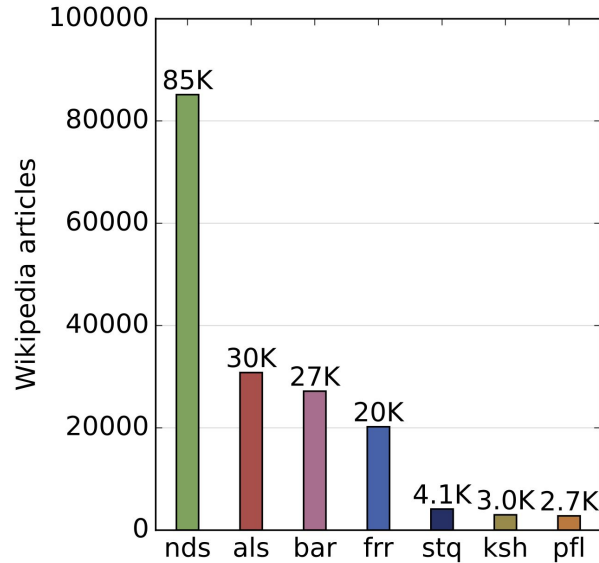
Saterfrisian (**stq**)

Riparian (**ksh**)

Rhine Franconian (**pfl**)



High lexical variation due to regional word choices and different pronunciations.

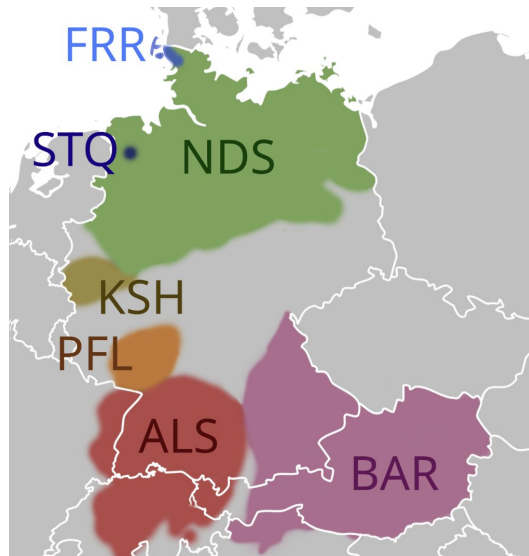
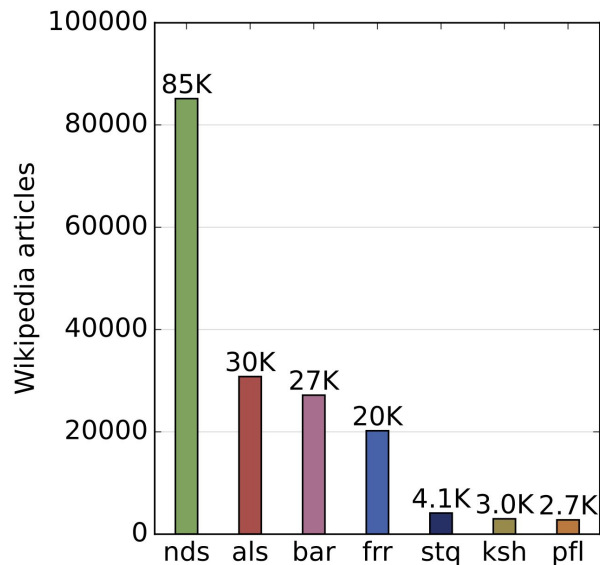


Low German (**nds**)
Alemannic (**als**)
Bavarian (**bar**)
North Frisian (**frr**)
Saterfrisian (**stq**)
Riparian (**ksh**)
Rhine Franconian (**pfl**)

Standard German: 2.9M Wiki articles



High lexical variation due to regional word choices and different pronunciations.



Low German (**nds**)
Alemannic (**als**)
Bavarian (**bar**)
North Frisian (**frr**)
Saterfrisian (**stq**)
Riparian (**ksh**)
Rhine Franconian (**pfl**)

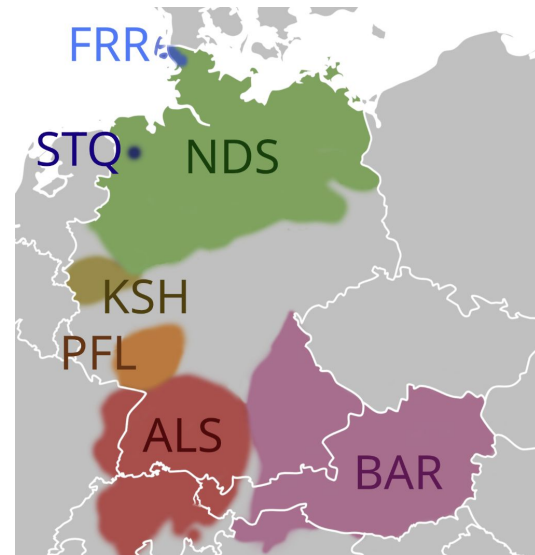
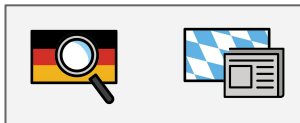


Low-resource: Limited resources available to train neural retrieval models.

Contribution

- New task: **Cross-dialect information retrieval**
- New dataset: **WikiDIR**
- Dialect variation **dictionaries**
- Evaluation of IR models on WikiDIR

Example



Agenda

1. **Motivation**
2. WikiDIR dataset
3. Dialect dictionaries
4. Models
5. Results

Agenda

1. Motivation
- 2. WikiDIR dataset**
3. Dialect dictionaries
4. Models
5. Results

Dataset Pipeline

Wikipedia
Dump of
dialect Y

Minga

190 Sproochen

Artikl dischkrian Leesen Werkeln Am Gwëntext werkeln Gschicht ähschaun Sunstigs

Der Artikl is im Dialekt **Mingarisch** gschriem worn.

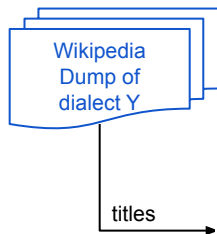
Fia andane Bedeitunga schau: **Minga (Begriffsklearung)**.

Minga (amtlī: **München**) Aussproch: [ˈmʏŋ(ː)ɐ] is d'Haptstod vo **Bayern**. In da Umgebung (20–30 km) hoaßt ma s'Minga oda oft aa oafach *d'Stod*. In Minga sogt ma München. Minga is mid mehra wia 1,5 Milliona Eihwohna d'gresste Stod vo Bayern und hinta **Berlin** und **Hamburg** d'drittgresste Stod vo **Deutschland**.^[2] D'Stod g'head zua d'wichtigstn Wirtschafts-, Vakeas- und Kuituazentren vo **Eiropa**. Minga is aa da Vawoitungssitz vom Regiarungsbeziak **Owabayern**.

Minga is in da ganzen woid aa zwengs da **Wiesen** und am **Hofbraihaus** bekannt. Dazua hods no vui andane Sengswiadigkeitm wias Glocknspui am Rathaus am **Marienplotz**, d'Residenz und s'**Schloss Nymphenburg**. Z'Minga gib't 's aa an Hauffa Museen, wias **Deitsche Museum**, oda d'**oide**, d'**neie** und d'**Pinakothek vo da Modeane**.

Woppn	Deitschlandkoatn
	
Basisdotn	
Bundesland:	Bayern
Regiarungsbeziak:	Owabayern

Dataset Pipeline



Query q_i



Minga 190 Sproochen

Leesen Warkeln Am Gwëntext warkeln Gschicht ähschaun Sunstigs

Der Articl is im Dialekt **Mingarisch** gschriem worn.

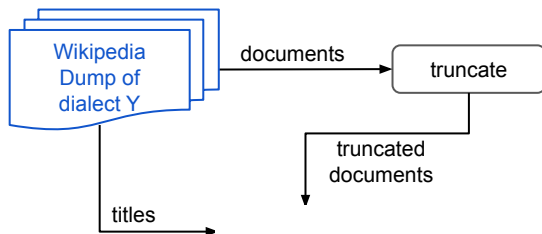
Fia andane Bedeitunga schau: **Minga (Begriffsklearung)**.

Minga (amtlil: **München**) Aussproch: [ˈmʏŋ(ː)ɐ] is d'Haptstod vo **Bayern**. In da Umgebung (20–30 km) hoaßt ma s'Minga oda oft aa oafach *d'Stod*. In Minga sogt ma München. Minga is mid mehra wia 1,5 Milliona Eihwohna d'gresste Stod vo Bayern und hinta **Berlin** und **Hamburg** d'drittgresste Stod vo **Deutschland**.^[2] D'Stod g'head zua d'wichtigstn Wirtschafts-, Vakeas- und Kuituazentren vo **Eiropa**. Minga is aa da Vawoitungssitz vom Regiarungsbeziak **Owabayern**.

Minga is in da ganzen woid aa zwengs da **Wiesen** und am **Hofbraihaus** bekannt. Dazua hods no vui andane Sengswiadigkeitm wias Glocknspui am Rathaus am **Marienplotz**, d'Residenz und s'**Schloss Nymphenburg**. Z'Minga gib't 's aa an Hauffa Museen, wias **Deutsche Museum**, oda d'**oide**, d'**neie** und d'**Pinakothek vo da Modeane**.

Woppn	Deitschlandkoatn
	
Basisdotn	
Bundesland:	Bayern
Regiarungsbeziak:	Owabayern

Dataset Pipeline



Query q_i



Corpus \mathcal{D}



Minga 190 Sproochen

Artikl dischkrian Leesen Werkeln Am Gwëntext werkeln Gschicht ähschaun Sunstigs

Der Artikl is im Dialekt **Mingarisch** gschriem worn.

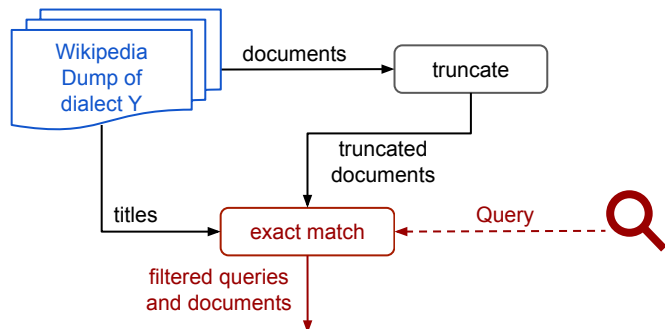
[Fia andere Bedeitunge schau: Minga \(Begriffsklearung\)](#)

Minga (amtl.: **München**) Aussproch: [ˈmʏŋ(ː)ɐ] is d'Haptstod vo [Bayern](#). In da Umgebung (20–30 km) hoaßt ma s'Minga oda oft aa oafach *d'Stod*. In Minga sogt ma München. Minga is mid mehra wia 1,5 Milliona Eihwohna d'gresste Stod vo Bayern und hinta [Berlin](#) und [Hamburg](#) d'drittgresste Stod vo [Deutschland](#).^[2] D'Stod g'head zua d'wichtigstn Wirtschafts-, Vakeas- und Kuituazentren vo [Europa](#). Minga is aa da Vawoitungssitz vom Regiarungsbeziak [Owabayern](#).

Minga is in da ganzen woid aa zwengs da [Wiesen](#) und am [Hofbraihaus](#) bekannt. Dazua hods no vui andere Sengswiadigkeitm wias Glocknspui am Rathaus am [Marienplotz](#), d'Residenz und s'[Schloss Nymphenburg](#). Z'Minga gib't 's aa an Hauffa Museen, wias [Deitsche Museum](#), oda d'*oide*, d'*neie* und d'[Pinakothek vo da Modeane](#).

Woppn	Deitschlandkoatn
	
Basisdotn	
Bundesland:	Bayern
Regiarungsbeziak:	Owabayern

Dataset Pipeline



Query q_i

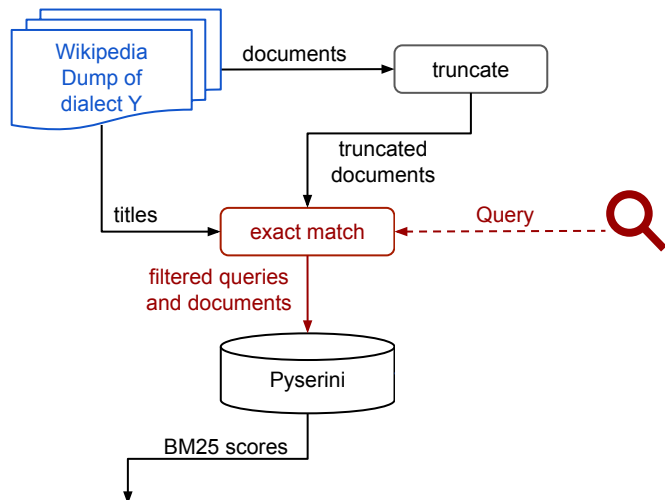


Corpus \mathcal{D}



$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

Dataset Pipeline



Query q_i



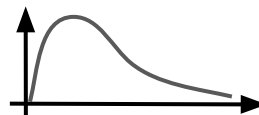
Corpus \mathcal{D}



$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

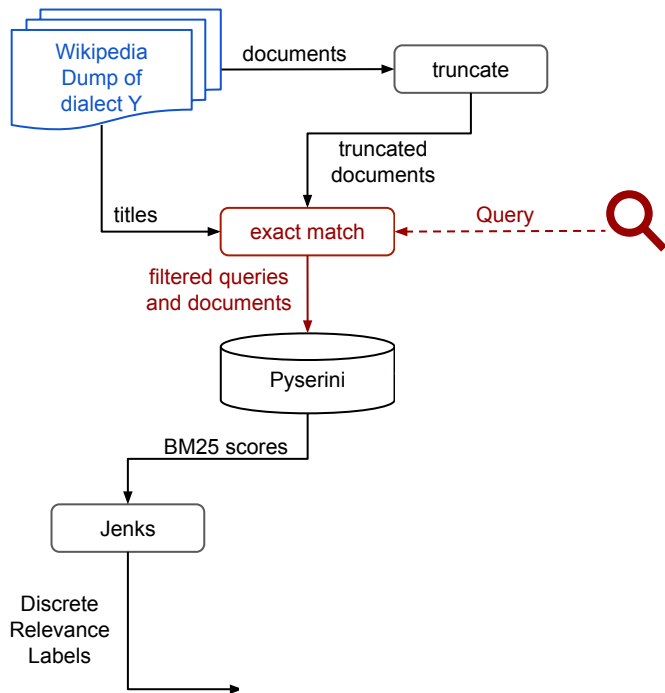


Lexical Similarity Scores



all (q,d)-pairs

Dataset Pipeline



Query q_i



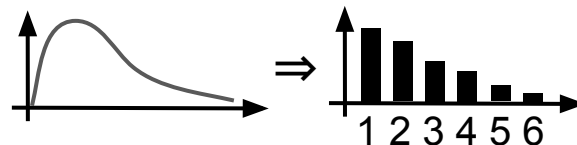
Corpus \mathcal{D}



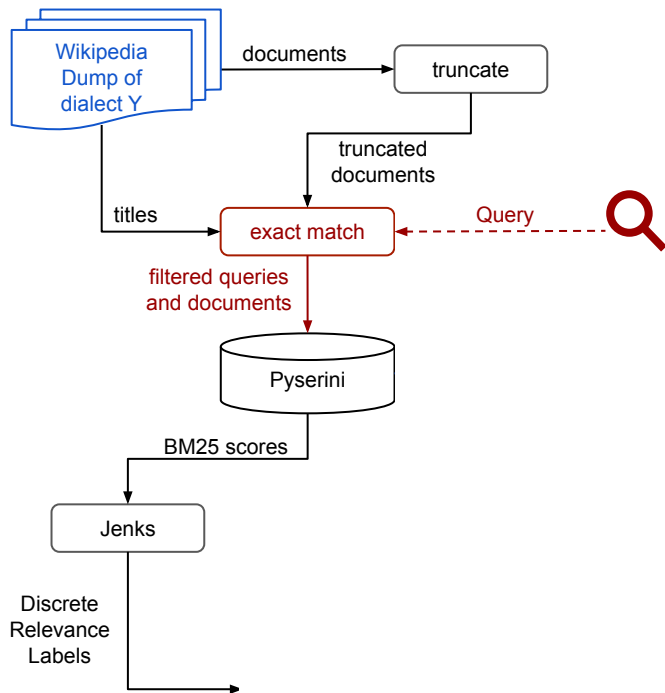
$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$



Monolingual Relevance Labels



Dataset Pipeline



Query q_i



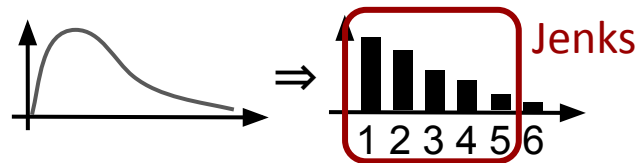
Corpus \mathcal{D}



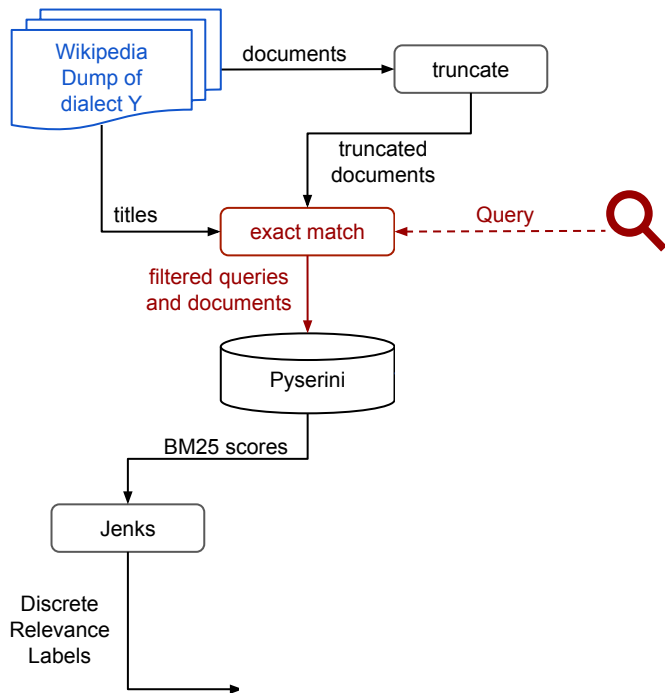
$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$



Monolingual Relevance Labels



Dataset Pipeline



Query q_i



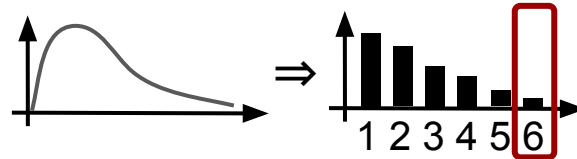
Corpus \mathcal{D}



$$\mathcal{D}_{\text{rel}}^{q_i} = \{d_j \in \mathcal{D} \mid d_j \text{ contains } q_i\}$$

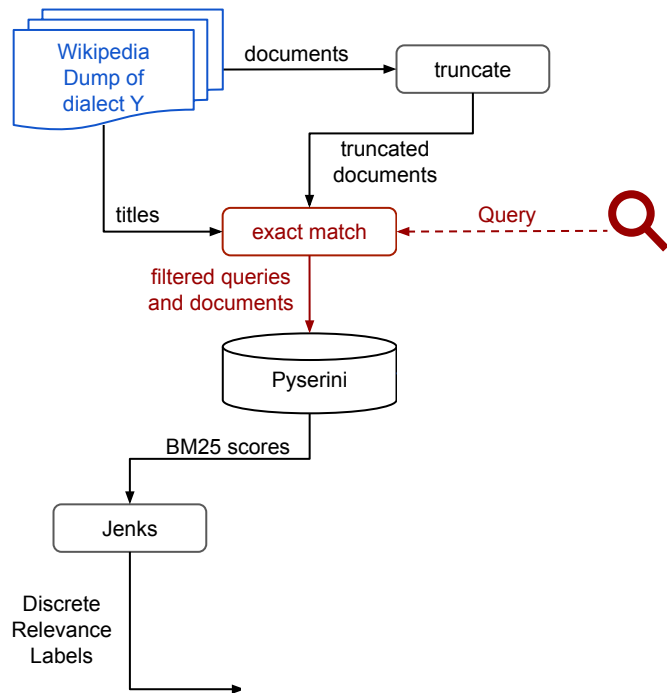


Monolingual Relevance Labels



same
article

Dataset Pipeline



Query q_i



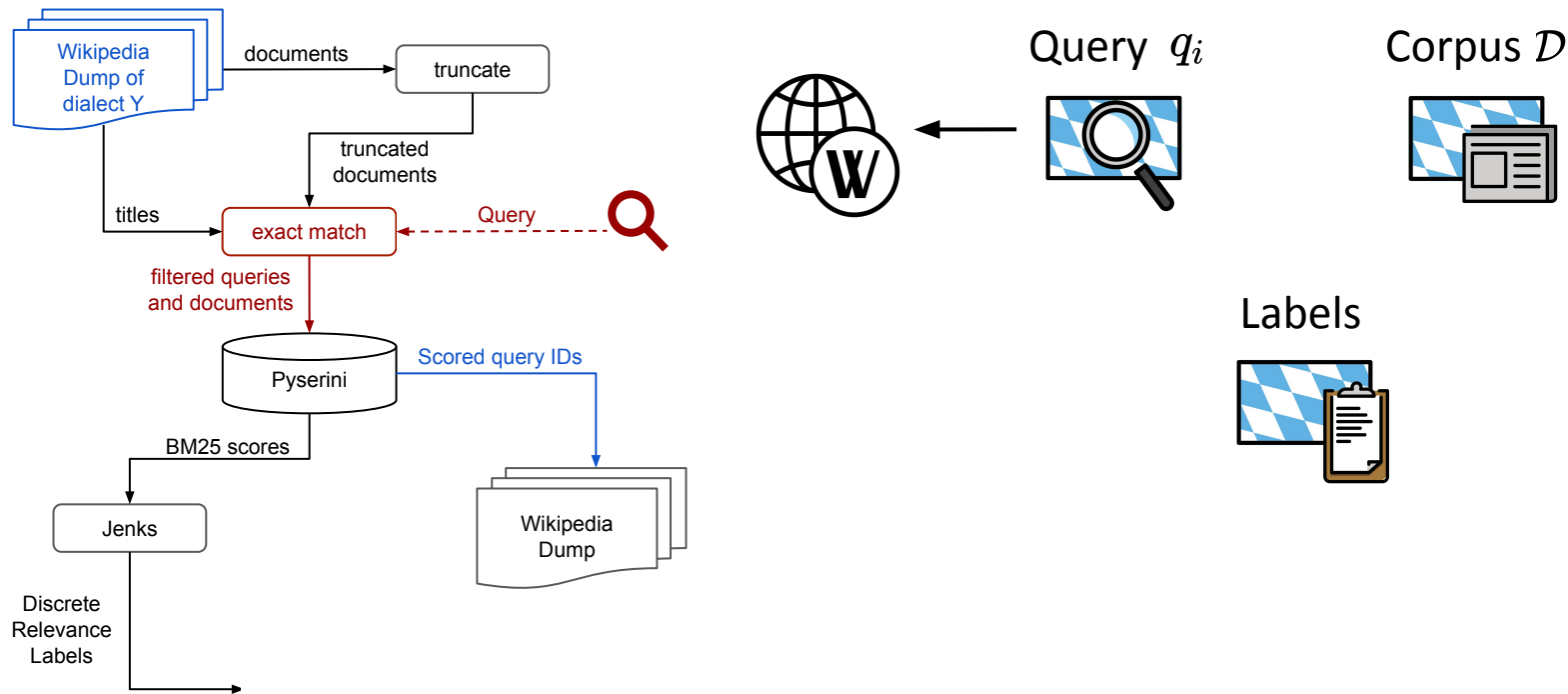
Corpus \mathcal{D}



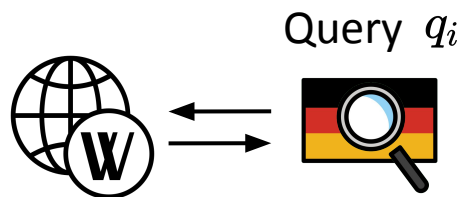
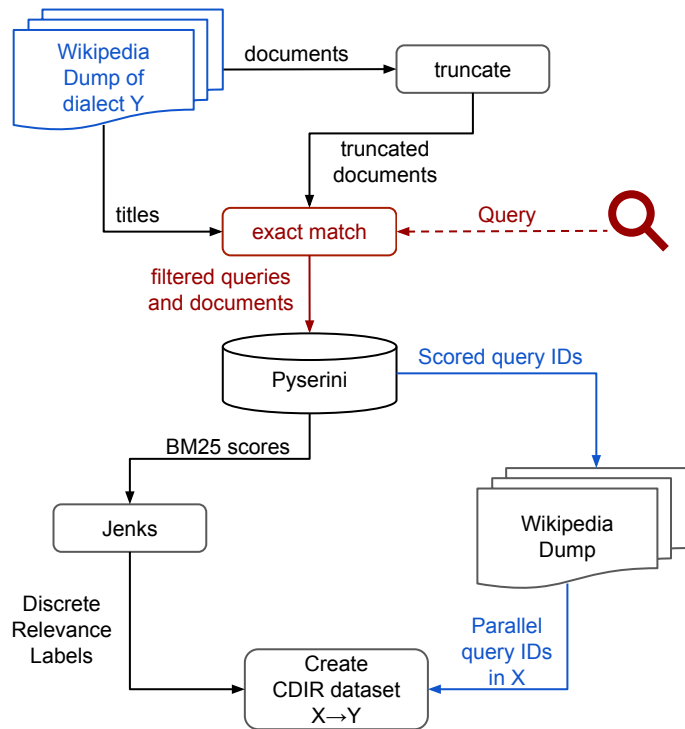
Labels



Dataset Pipeline



Dataset Pipeline



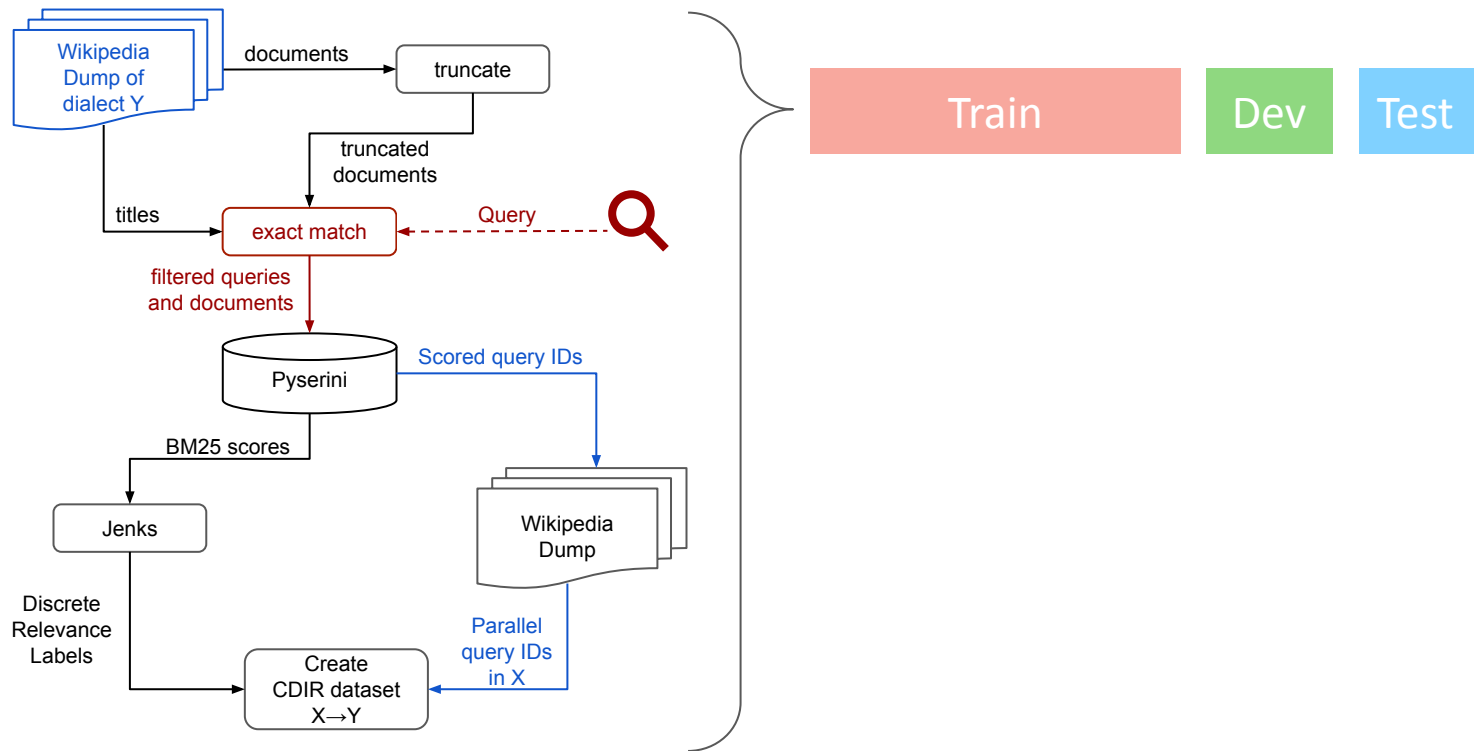
Labels



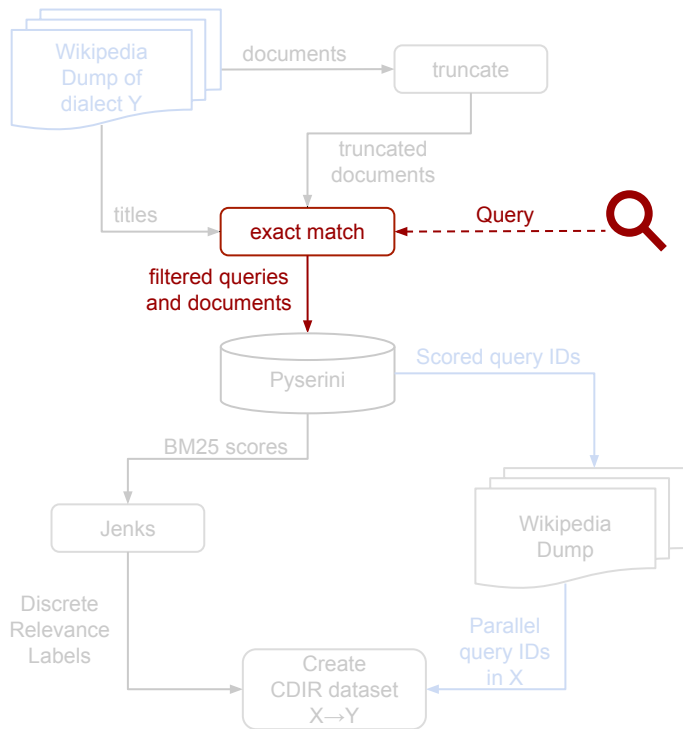
Corpus \mathcal{D}



Dataset Pipeline



Dataset Pipeline



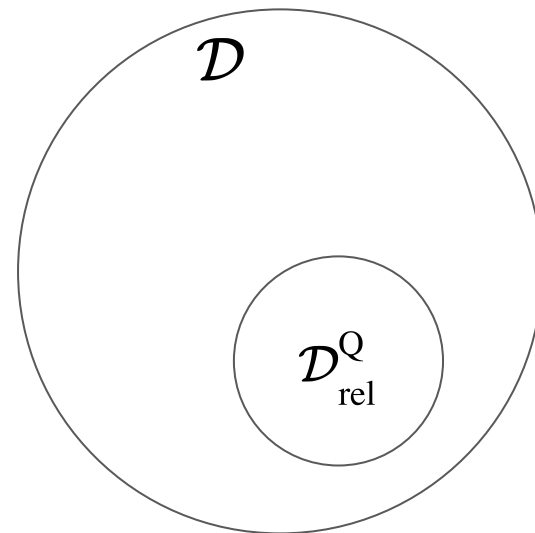
Train

Dev

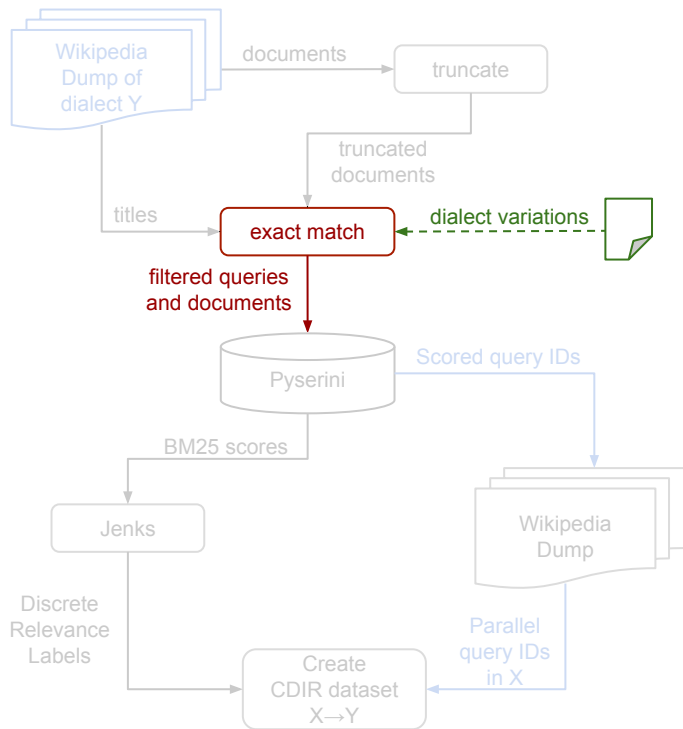
Test

Set of rel. docs.

All documents
that contain a
query.

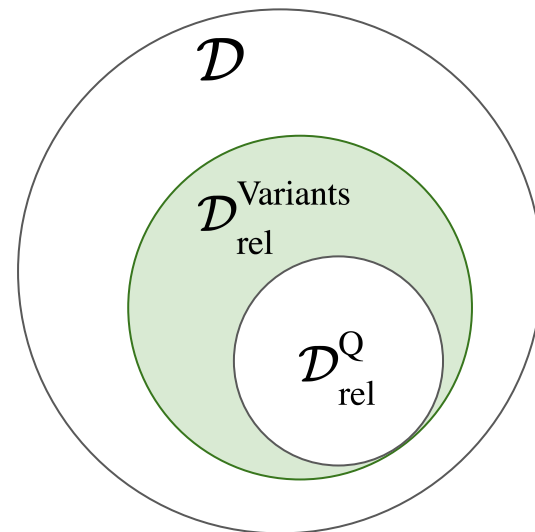


Dataset Pipeline

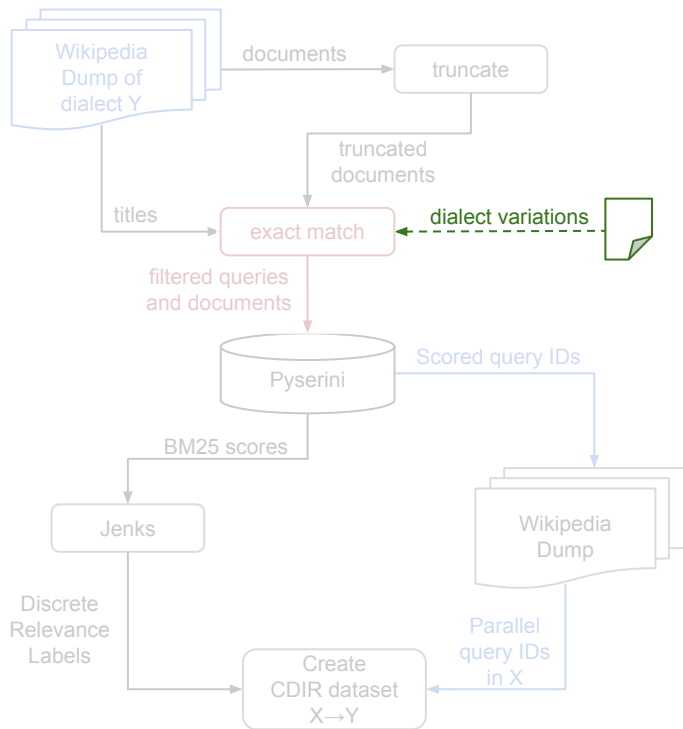


Analysis Split

All documents that contain a query or any of its dialect variations.



Dataset Pipeline

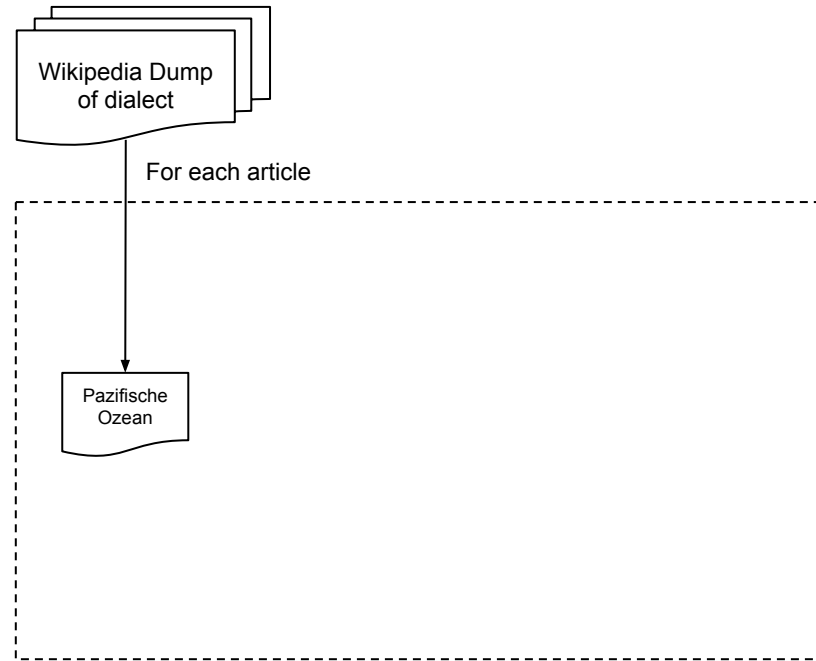


Where do **dialect variations** come from?

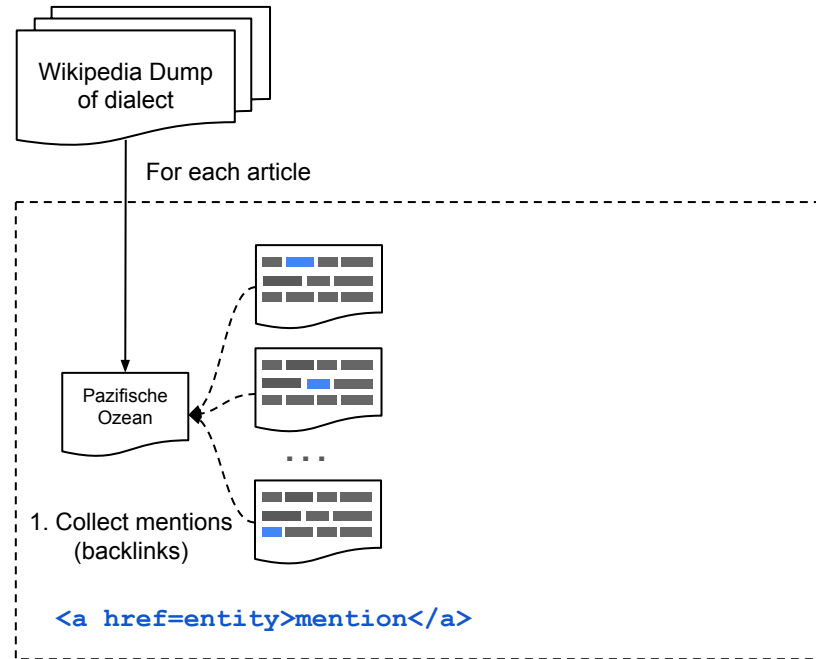
Agenda

1. Motivation
2. WikiDIR Dataset
- 3. Dialect dictionaries**
4. Models
5. Results

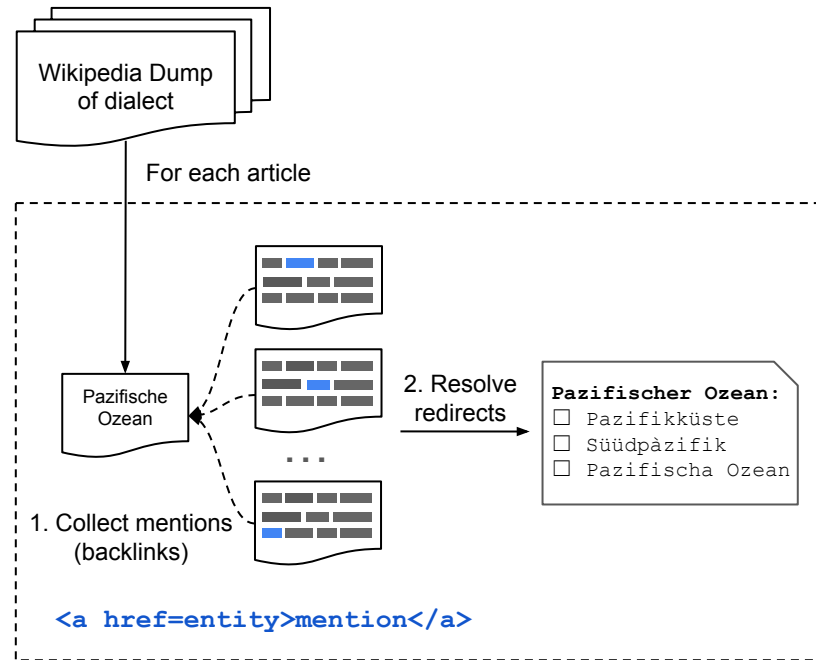
Dialect variation dictionaries



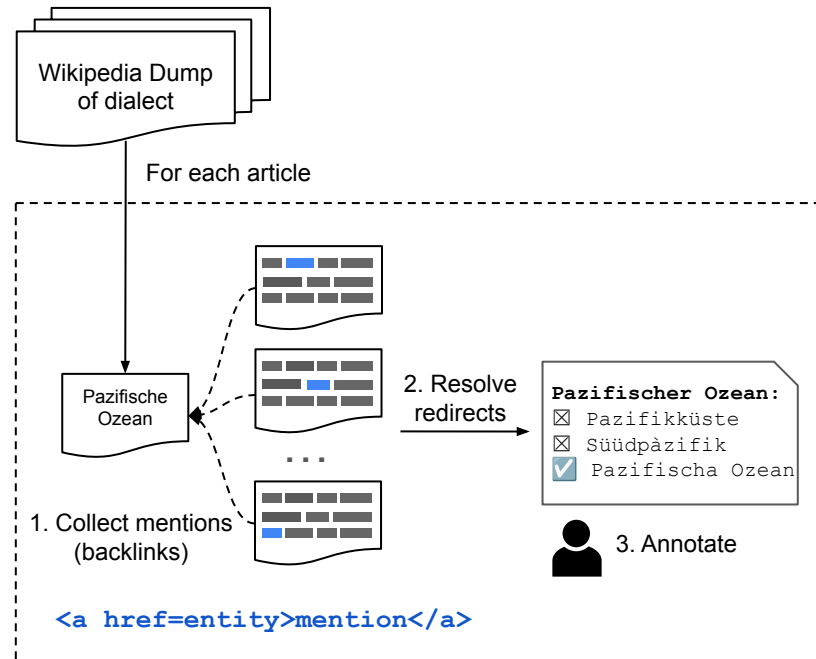
Dialect variation dictionaries



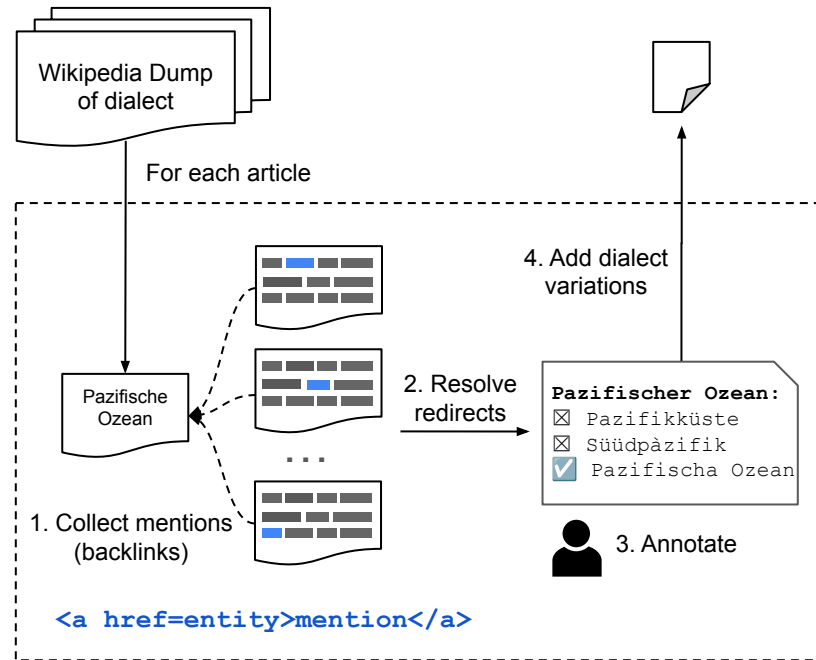
Dialect variation dictionaries



Dialect variation dictionaries



Dialect variation dictionaries



Example Record (Bavarian dictionary)

```
{  
  "de_id": "3215",  
  "de_title": "München",  
  "dial_id": "12259",  
  "dial_title": "Minga",  
  "variants": ["Müñch'n", "Minkcha", "Minkn", "Minchn", "Mingna", "Minkhn", "Münchn"]  
}
```

Agenda

1. Motivation
2. WikiDIR dataset
3. Dialect dictionaries
- 4. Models**
5. Results

Models

Baseline: BM25 (Robertson, 1995)

Models



RankGPT (Llama 3.1)

===== LLM-RERANKING =====

system: You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

user: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)

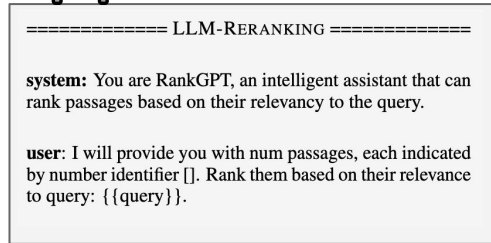
Baseline: BM25 (Robertson, 1995)

Llama icon created by Freepik - Flaticon

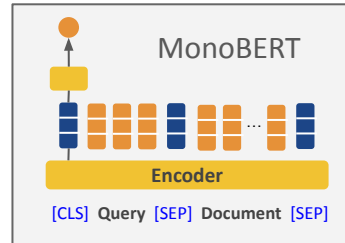
Models



RankGPT (Llama 3.1)



(Sun et al., 2023)



(Nogueira et al., 2019)

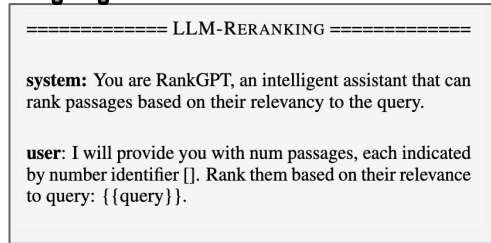
Baseline: BM25 (Robertson, 1995)

Model diagrams taken from (Litschko, 2024)

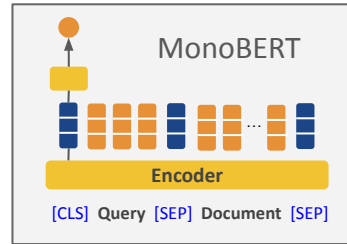
Models



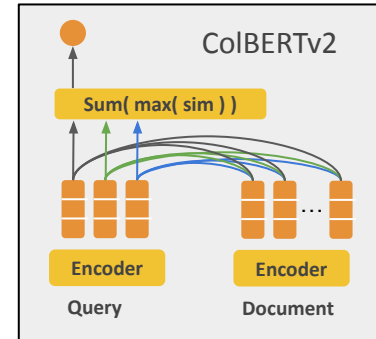
RankGPT (Llama 3.1)



(Sun et al., 2023)



(Nogueira et al., 2019)



(Santhanam et al., 2022)

Baseline: BM25 (Robertson, 1995)

Model diagrams taken from (Litschko, 2024)

Models

Rerank top 100



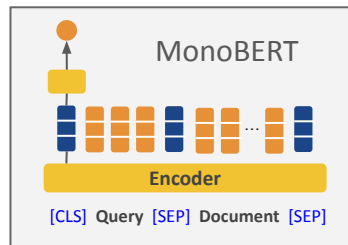
RankGPT (Llama 3.1)

===== LLM-RERANKING =====

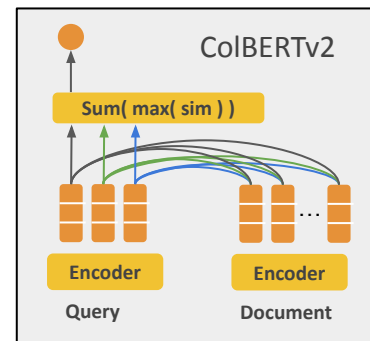
system: You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

user: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)



(Nogueira et al., 2019)



(Santhanam et al., 2022)

Baseline: BM25 (Robertson, 1995)

Models



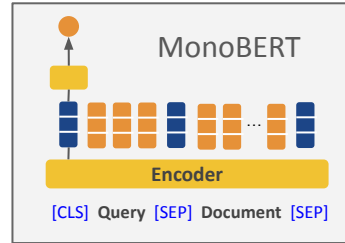
RankGPT (Llama 3.1)

===== LLM-RERANKING =====

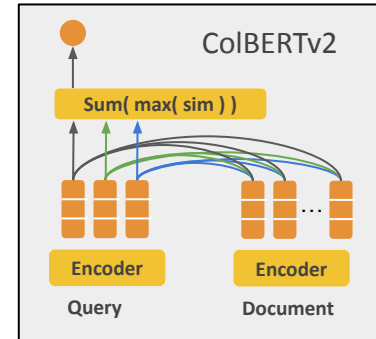
system: You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

user: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)



(Nogueira et al., 2019)



(Santhanam et al., 2022)

Baseline: BM25 (Robertson, 1995)

Models



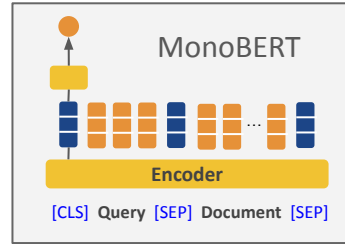
RankGPT (Llama 3.1)

===== LLM-RERANKING =====

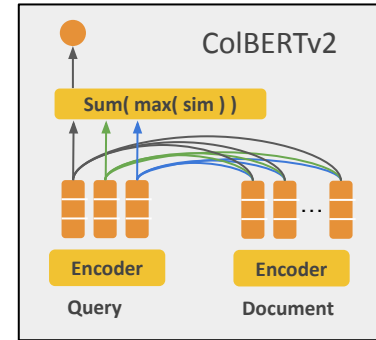
system: You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

user: I will provide you with num passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

(Sun et al., 2023)

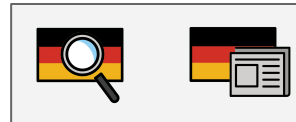


(Nogueira et al., 2019)

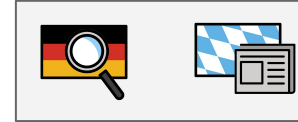


(Santhanam et al., 2022)

Zero-shot Transfer



Fine-tuning

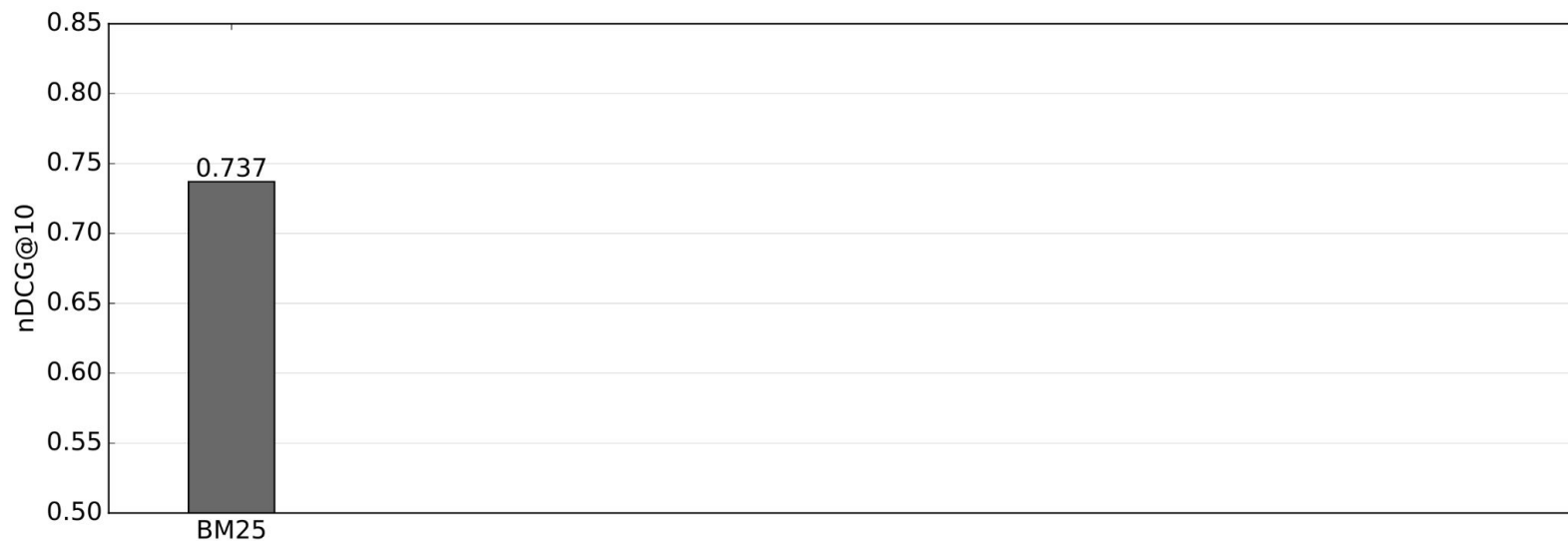


Baseline: BM25 (Robertson, 1995)

Agenda

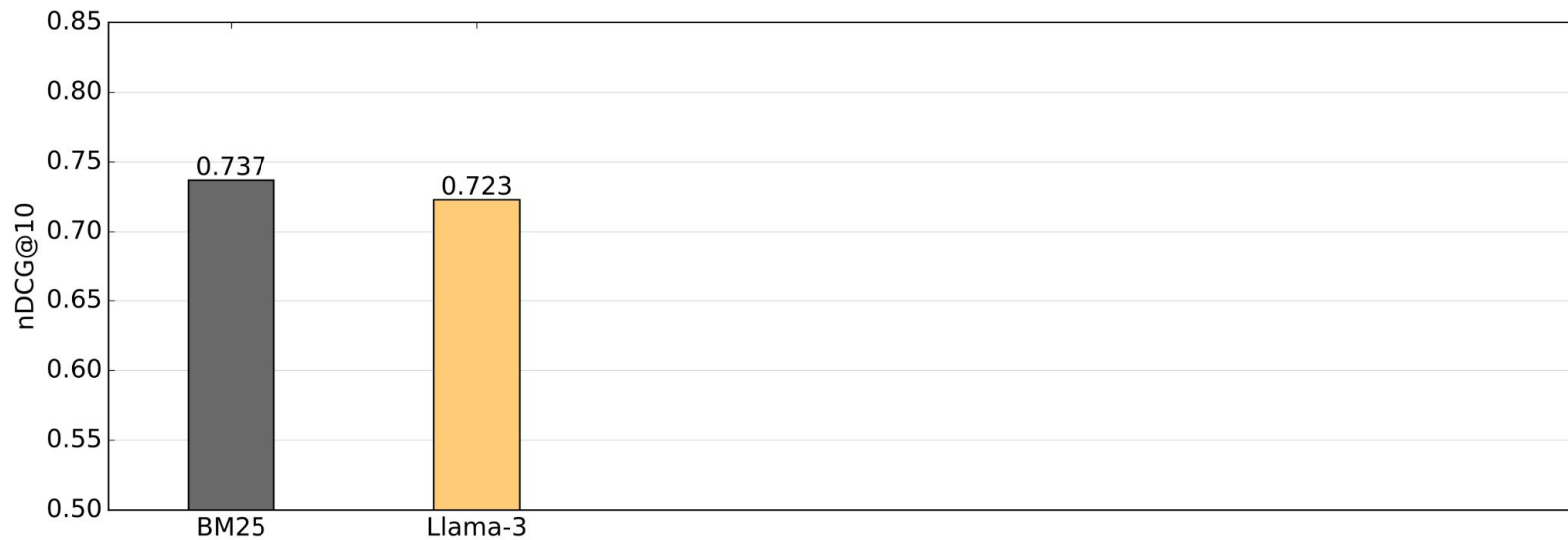
1. Motivation
2. WikiDIR dataset
3. Dialect dictionaries
4. Models
- 5. Results**

Main results



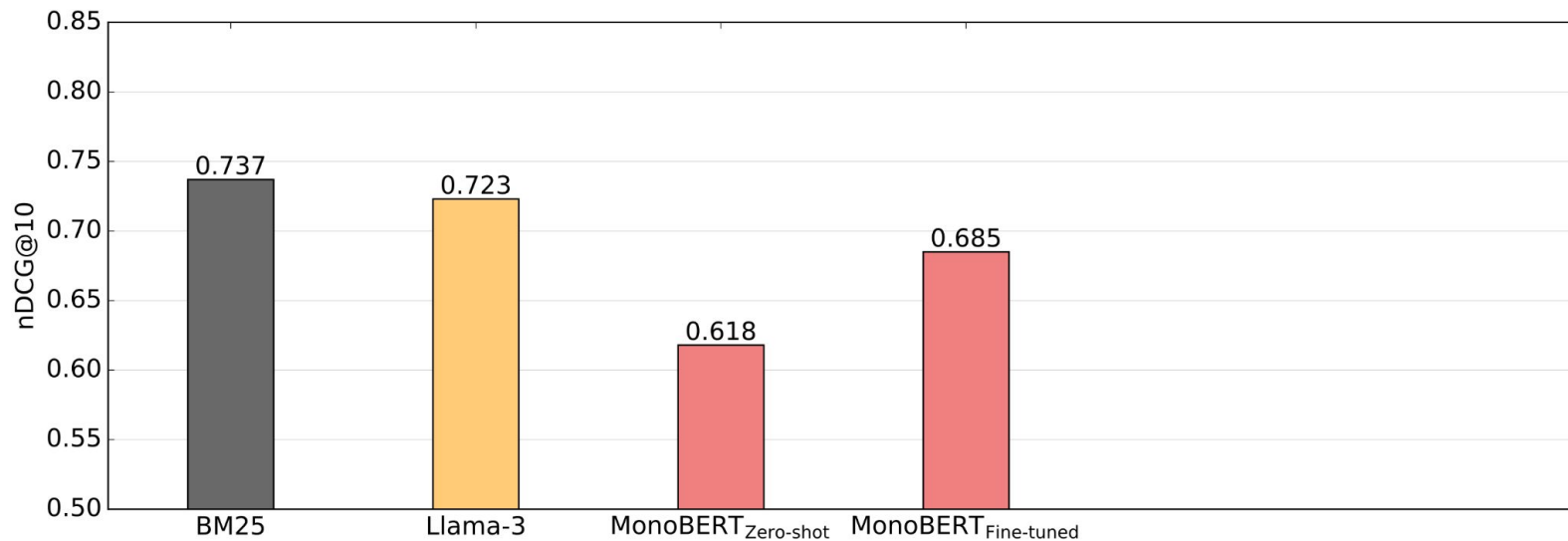
*average over 7 dialects

Main results



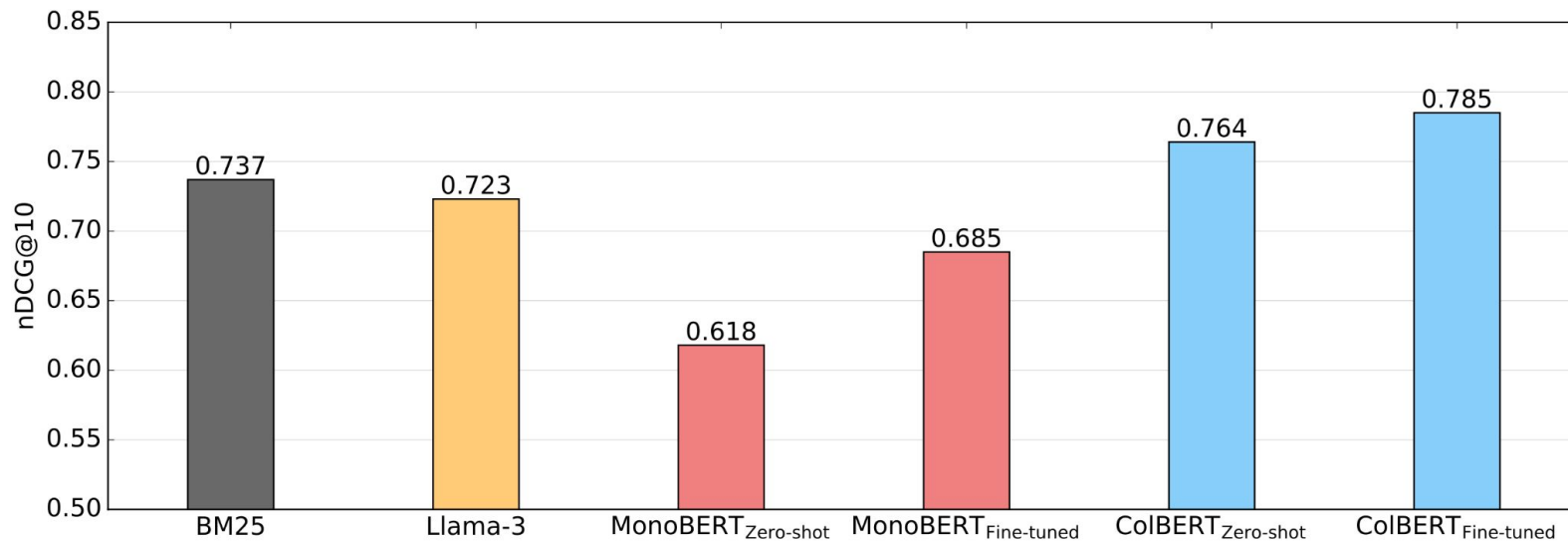
*average over 7 dialects

Main results



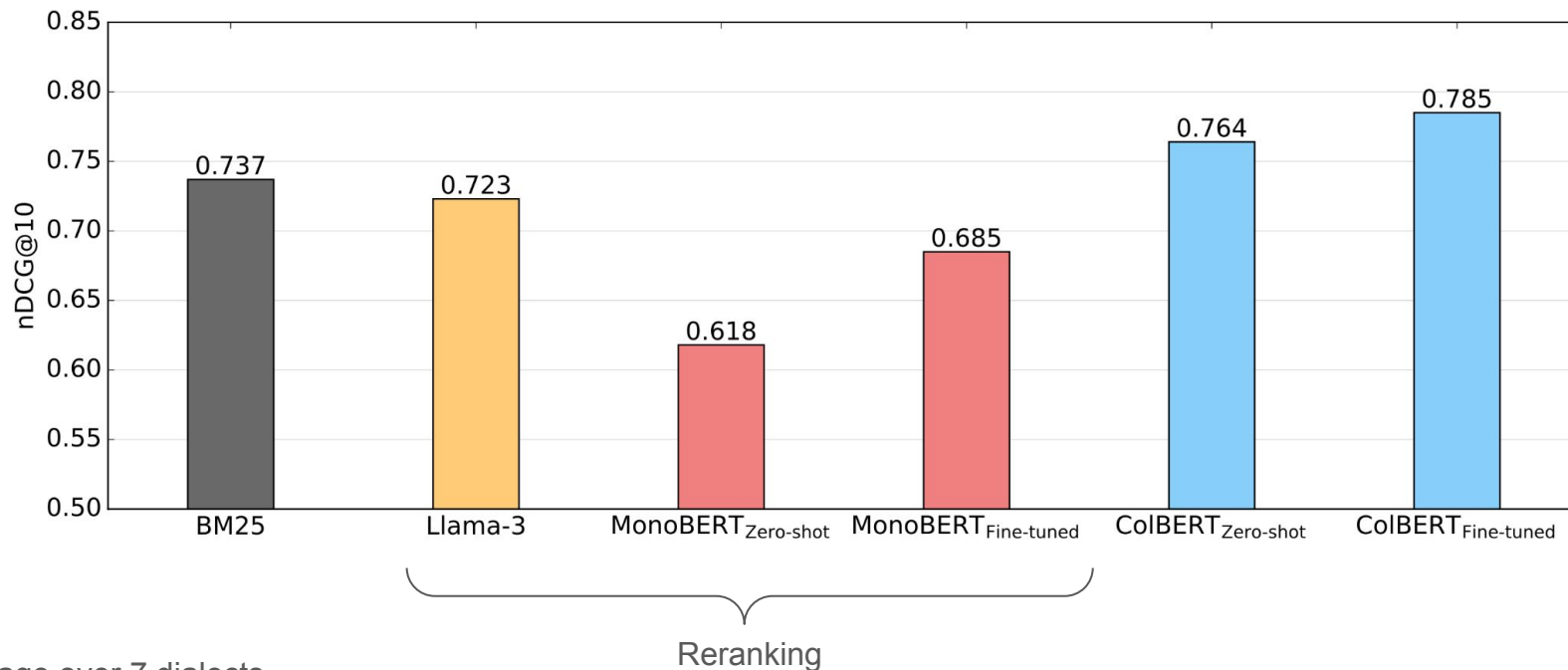
*average over 7 dialects

Main results



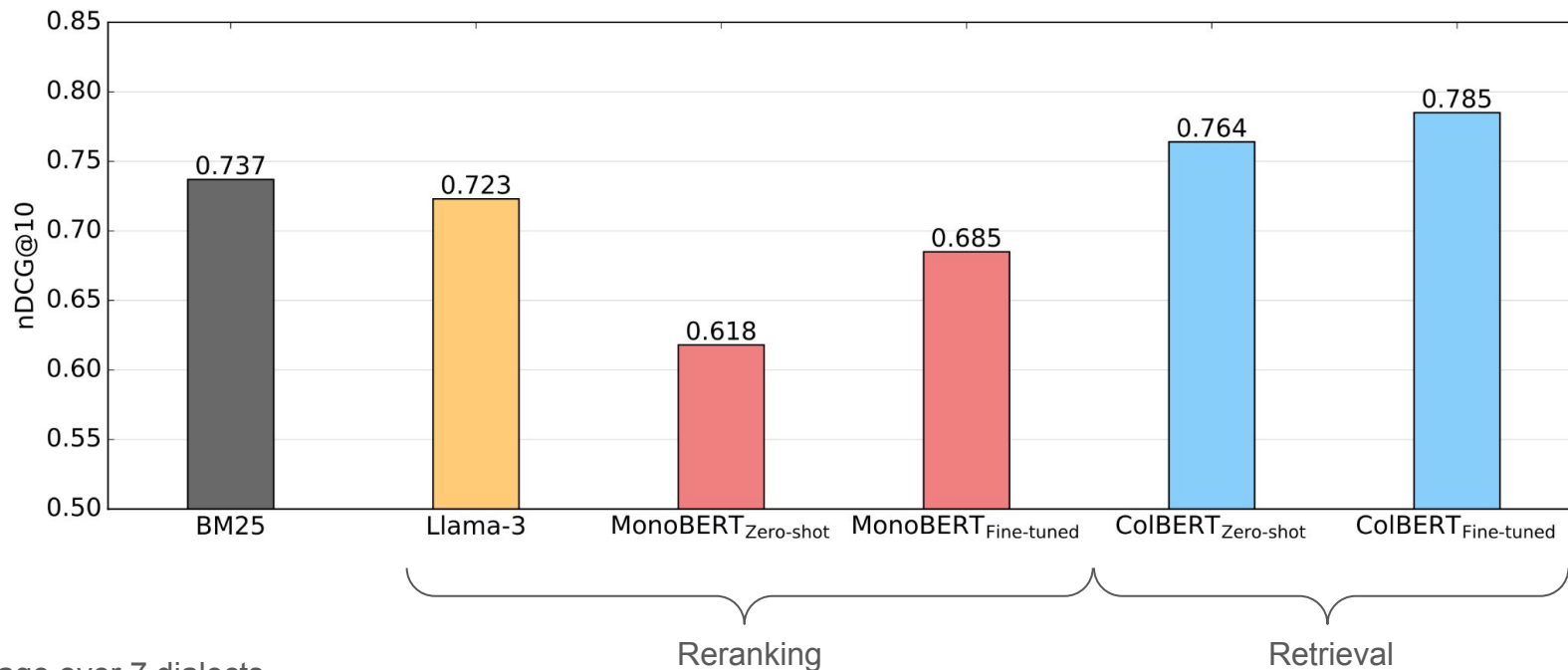
*average over 7 dialects

Main results



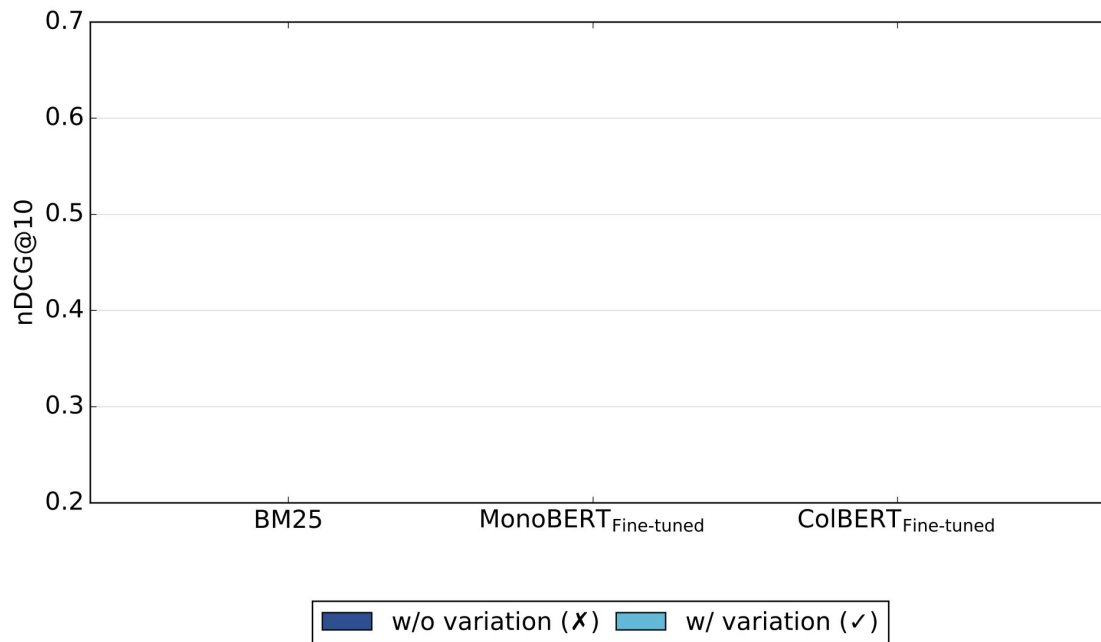
*average over 7 dialects

Main results



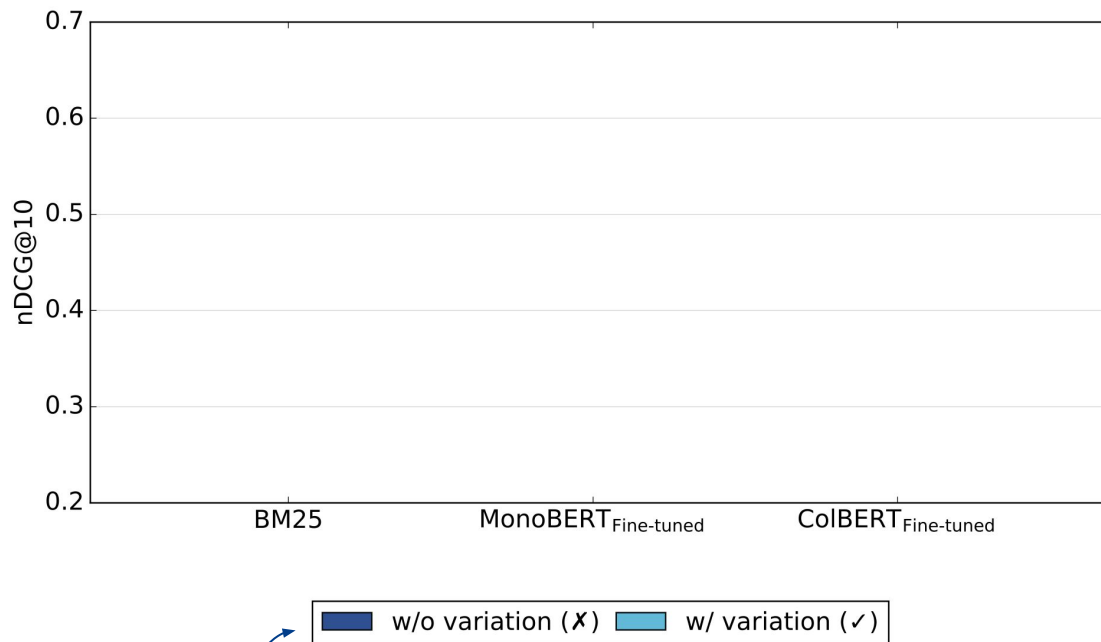
*average over 7 dialects

Dialect variation results



*average over 5 dialects

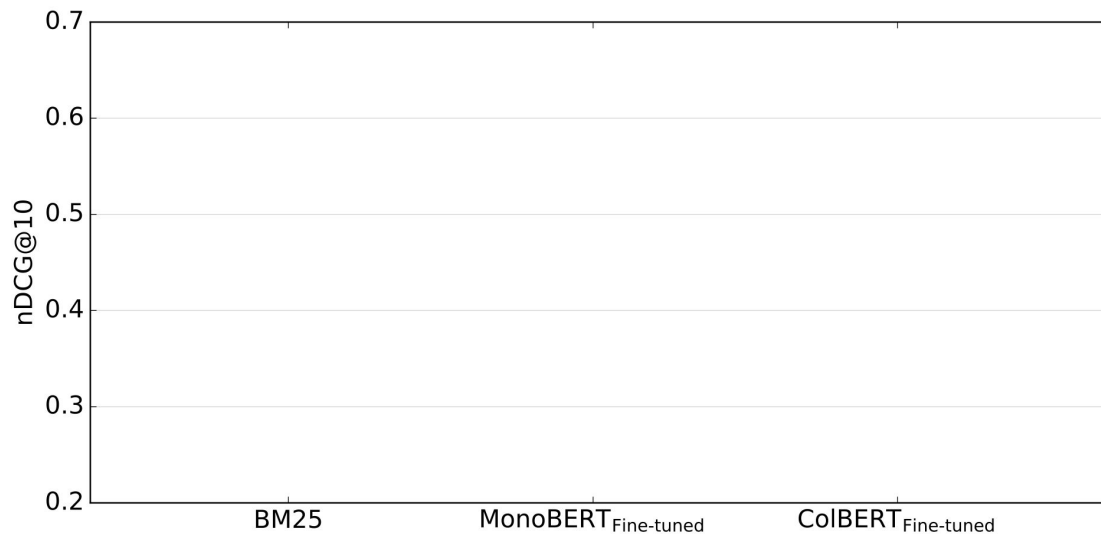
Dialect variation results



*average over 5 dialects

Exclude documents
containing variations

Dialect variation results

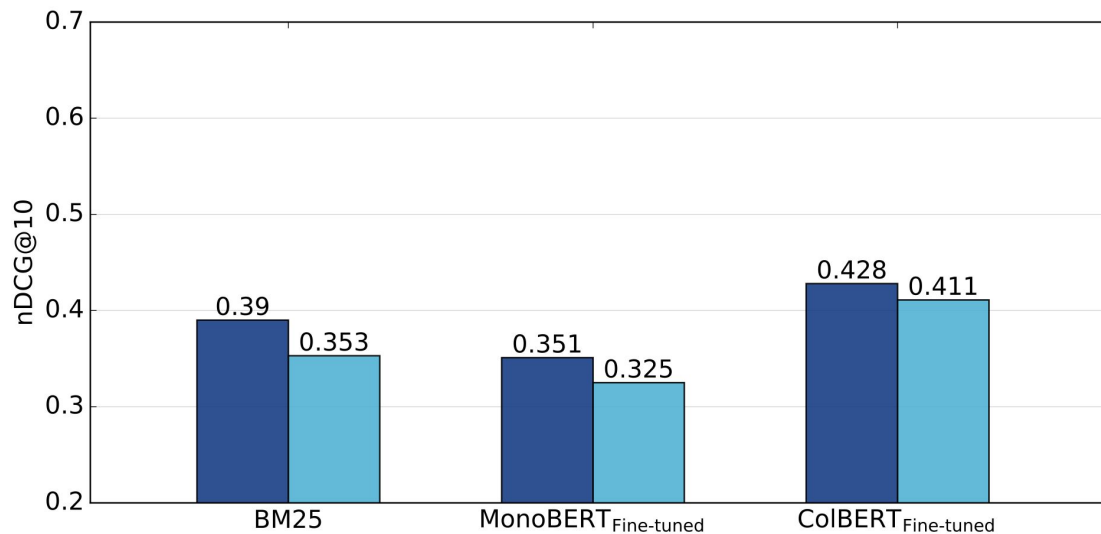


*average over 5 dialects

Exclude documents
containing variations

full analysis split

Dialect variation results



*average over 5 dialects

Exclude documents
containing variations

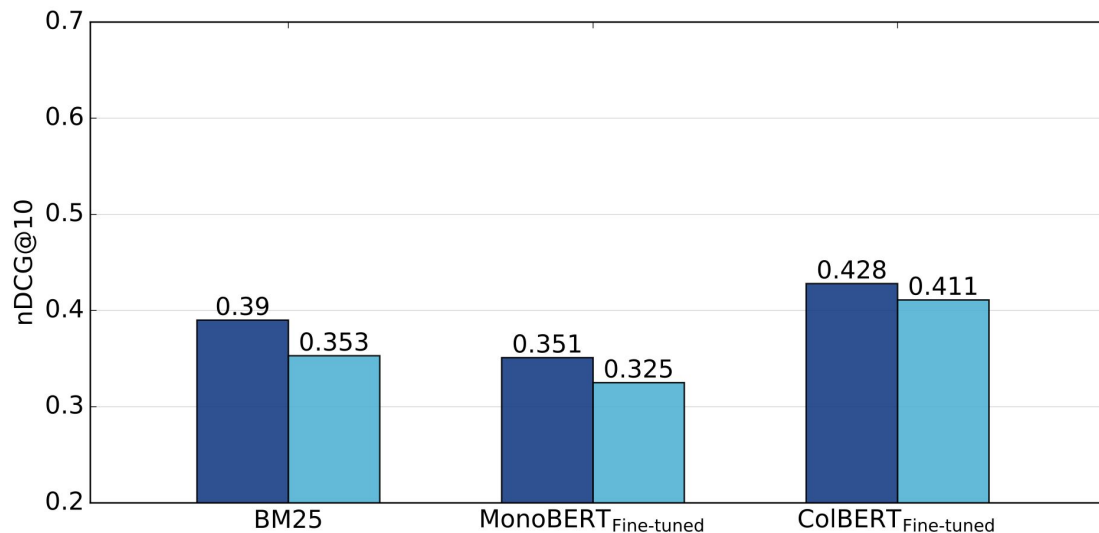
full analysis split

Document translation results

Can we use LLMs to close the dialect gap?



Document transl.
Dialect → DE



*average over 5 dialects

Exclude documents
containing variations

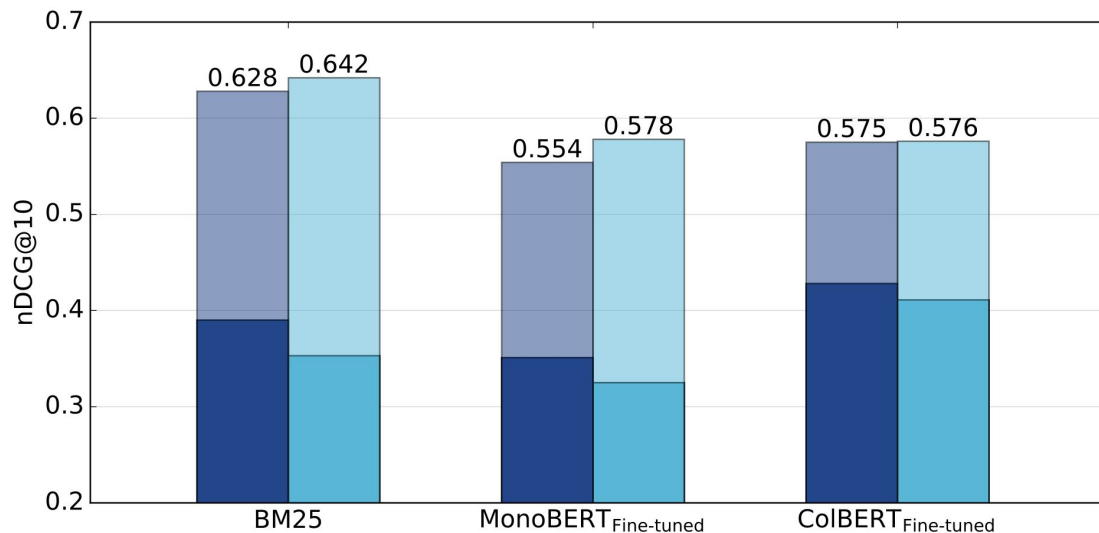
full analysis split

Document translation results

Can we use LLMs to close the dialect gap?



Document transl.
Dialect → DE



*average over 5 dialects

Exclude documents
containing variations

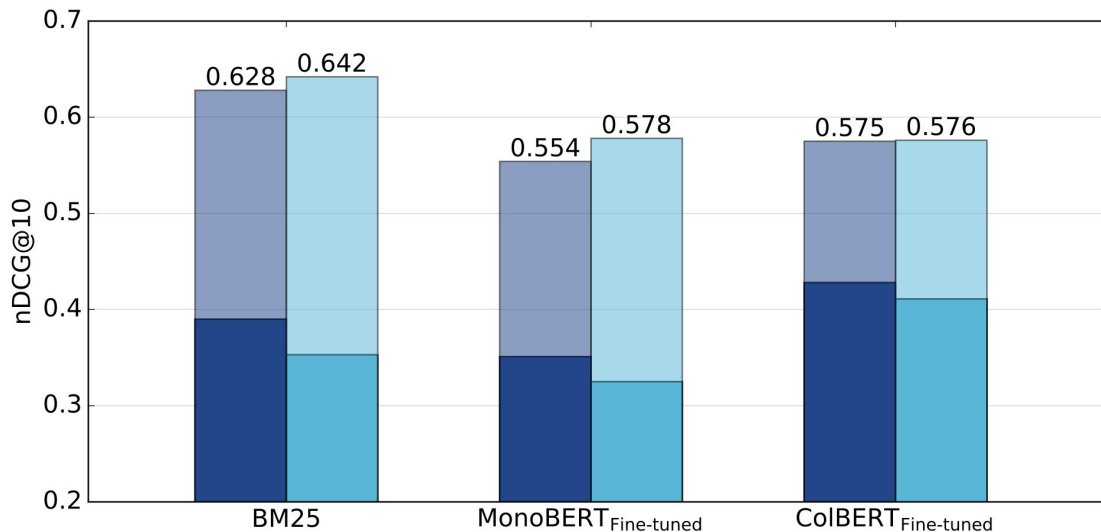
full analysis split

Document translation results

Can we use LLMs to close the dialect gap?



Document transl.
Dialect → DE



There are still large gaps!

*average over 5 dialects

Exclude documents containing variations

full analysis split

Conclusion

- We introduce [WikiDIR](#), a cross-dialect information retrieval dataset.
- We release [dialect variation dictionaries](#) for German dialects.
- More results and analyses in the paper.

GitHub



Conclusion

- We introduce [WikiDIR](#), a cross-dialect information retrieval dataset.
- We release [dialect variation dictionaries](#) for German dialects.
- More results and analyses in the paper.

CDIR is **novel and challenging** task!

→ Low-resource

→ High-Variance

The **gaps are still large**.

GitHub

