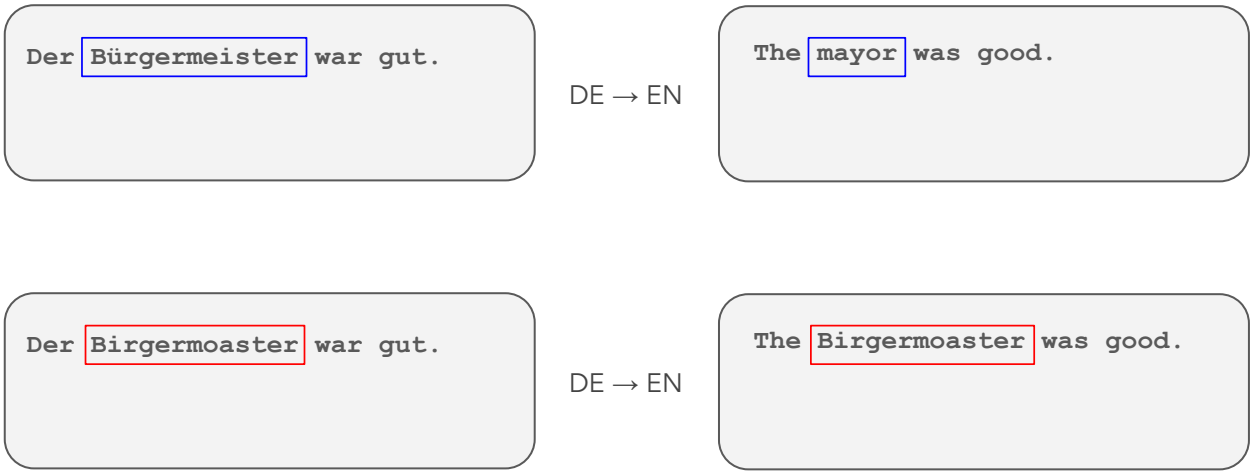
Make Every Letter Count:
Building Dialect Variation Dictionaries from Monolingual Corpora

Robert Litschko, Verena Blaschke, Diana Burkhardt, Barbara Plank, Diego Frassinelli

robert.litschko@lmu.de

Munich Center for Machine Learning

Machine Translation



→ Translation is not robust to regional spelling variation (Bavarian).

Wikipedia Search

(German)

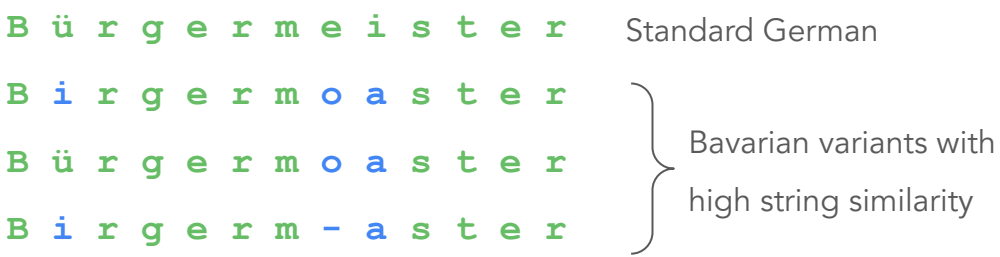
(Bavarian)

→ Preprocessing does not normalize Bavarian search query.

Source: bar.wikipedia.org

Motivation

Dialects are low-resource languages, existing tools are not robust to regional spelling variations.



Contributions

- DiaLemma: We present a novel framework to build dialect variation dictionaries.
- We build a dialect variation dictionary containing Bavarian translations for 5,124 German lemmas.
- Dialect NLP Tasks: We evaluate how well large language models (LLMs) recognize / translate dialect variants.

Agenda

Motivation

Annotation Framework

Dialect NLP Tasks

Results

Conclusion

Agenda

Motivation

Annotation Framework

Dialect NLP Tasks

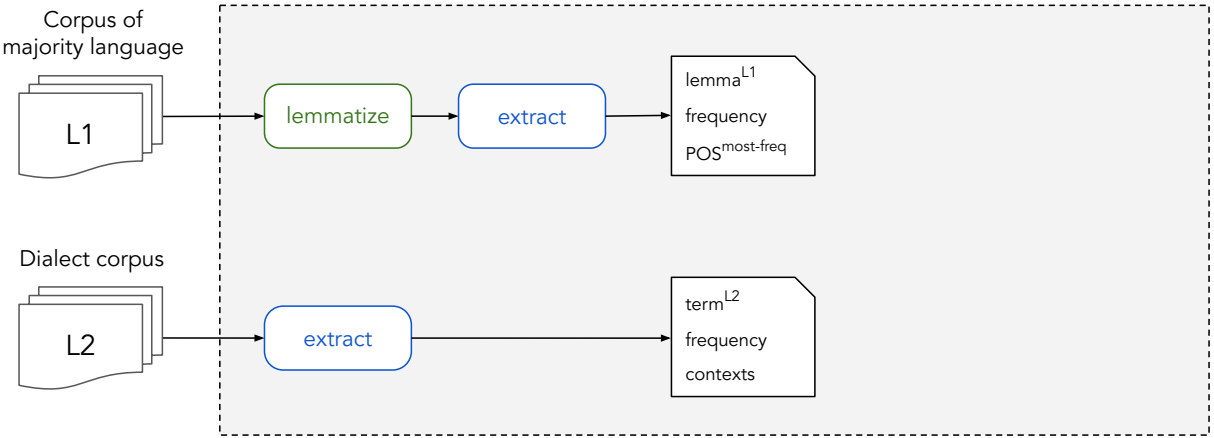
Results

Conclusion

DIALEMMA Annotation Framework

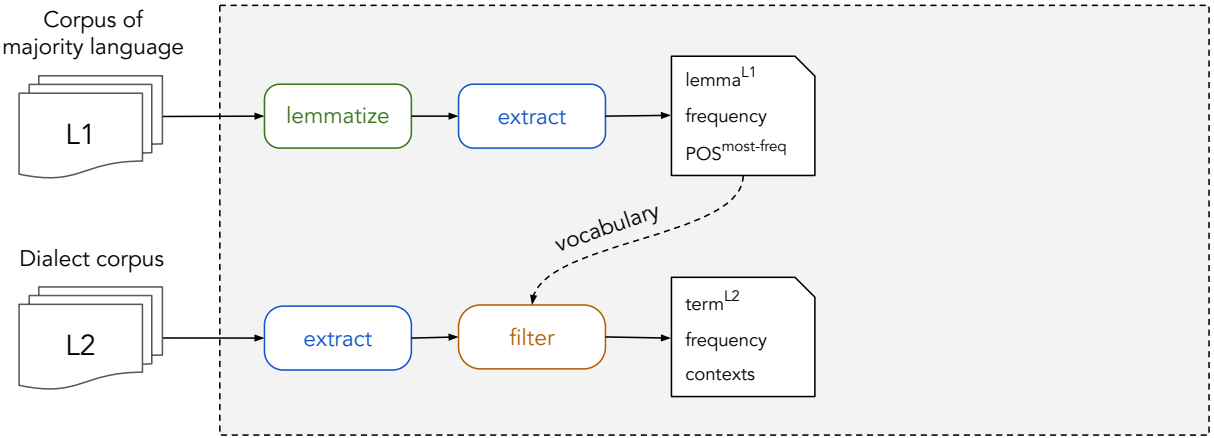
Step 1: Lemmatize the corpus of the majority language.

DIALEMMA Annotation Framework



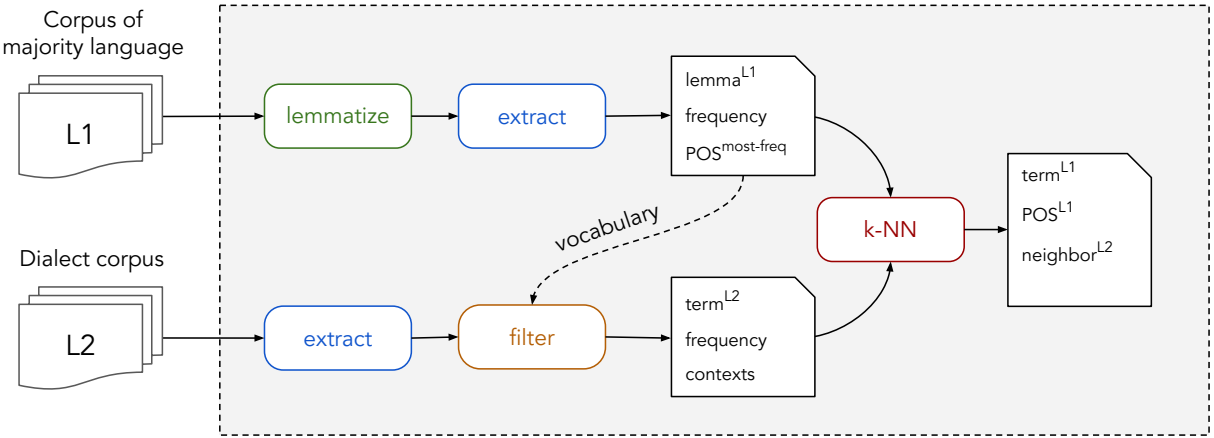
Step 2: Extract vocabularies from both corpora.

DIALEMMA Annotation Framework



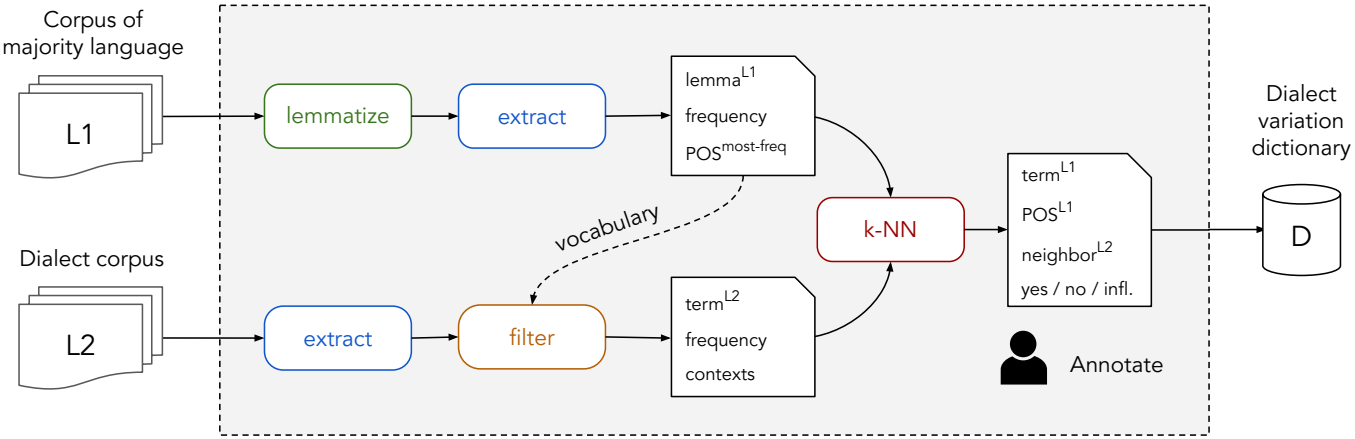
Step 3: Filter out lemmas (L1) from dialect vocabulary (L2).

DIALEMMA Annotation Framework



Step 4: For each lemma, extract lexically most similar dialect terms.

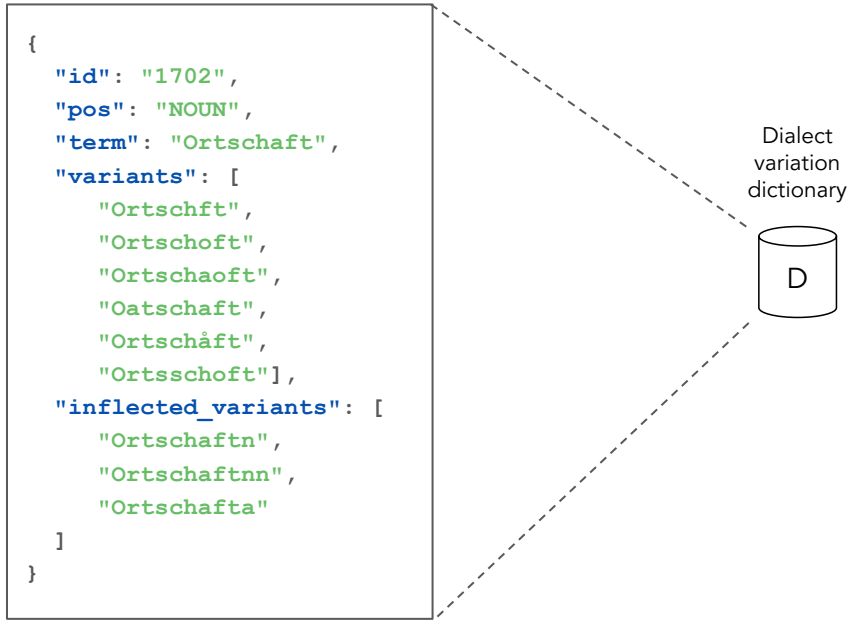
DIALEMMA Annotation Framework



Step 5: Annotate if dialect terms are direct / inflected translations.

Annotated Dataset

- We annotated 100K Bavarian German word pairs.
- 5,124 German lemmas.
  - 2.61 ± 1.88 dialect variants.
  - 2.57 ± 1.90 inflected variants.



Agenda

Motivation

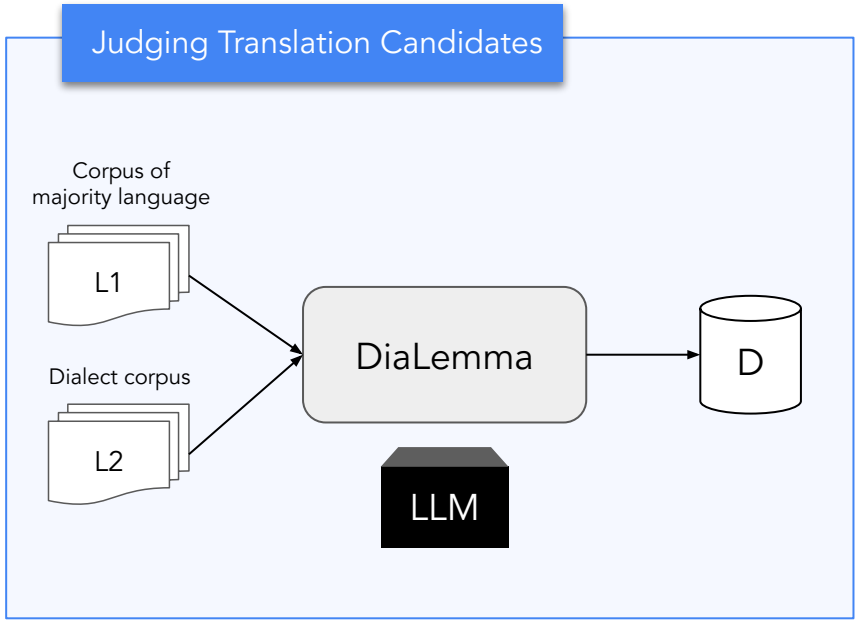
Annotation Framework

Dialect NLP Tasks

Results

Conclusion

Dialect NLP Tasks

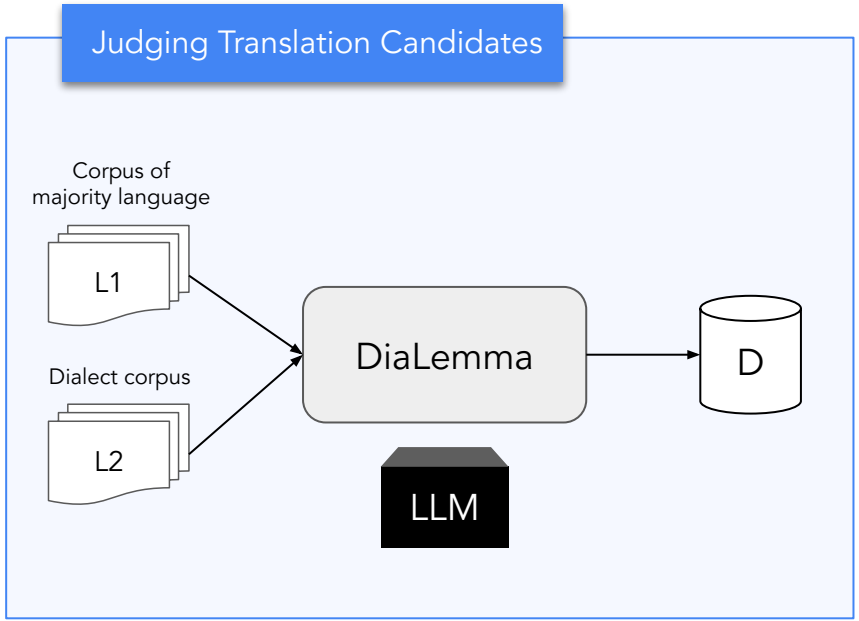


Three-way Word-Pair Classification Task

German Lemma	Bavarian Candidate	Translation?
zweisprachig (“billingual”)	zwaasprochig	yes
	zwaspråchig	yes
	zwoasprachign	inflected
	measprochig (“multilingual”)	no
	…	…

Performance is measured as the macro F1-score across all three classes.

Dialect NLP Tasks

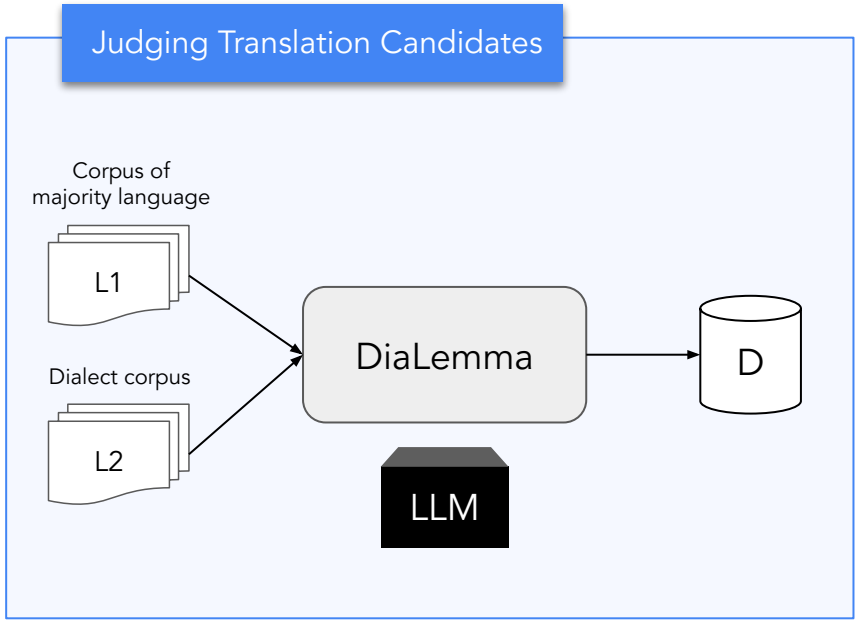


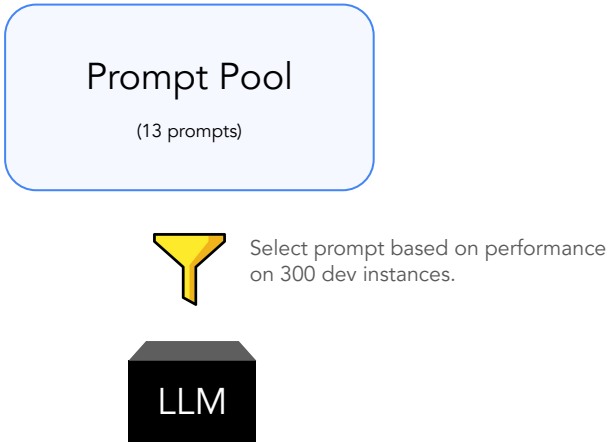
Three-way Word-Pair Classification Task

Part-of-Speech	Yes	Inflected	No
Noun	6,720	2,670	28,480
Adjective	1,358	3,066	5,496
Adverb	1,157	—	2,783
Verb	934	1,182	6,214
Proper Noun	574	86	34,430
Total	11,044	7,070	81,586

Testset with 97K instances across fifteen POS categories.

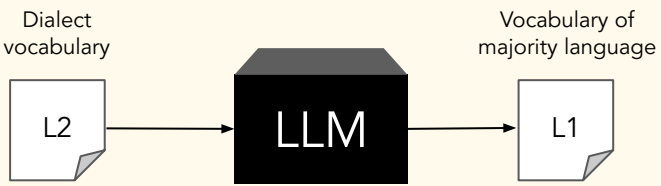
Dialect NLP Tasks





Dialect NLP Tasks

Dialect-to-Standard Translation



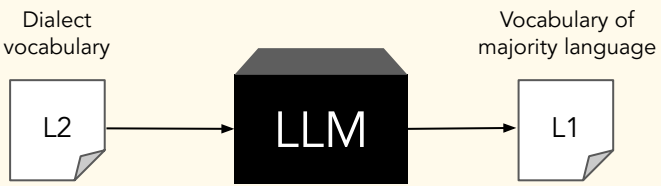
Word Translation Task

Bavarian (source)	German Lemma (target)
zwaasprochig	zweisprachig
zwaspråchig	zweisprachig
dozwischn	dazwischen
…	…

Performance is measured as the proportion of correct translations (translation accuracy).

Dialect NLP Tasks

Dialect-to-Standard Translation



Word Translation Task

Part-of-Speech	German Lemma (target)
Noun	6,564
Adjective	1,325
Adverb	1,126
Verb	916
Proper Noun	556
Total	10,775

Testset with 10.8K instances (300 dev set instances, pool: 21 prompts).

Agenda

Motivation

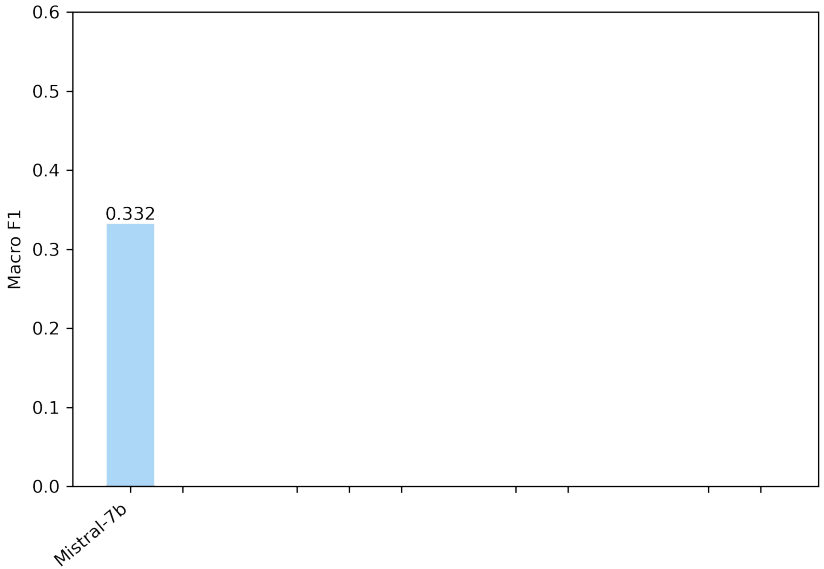
Annotation Framework

Dialect NLP Tasks

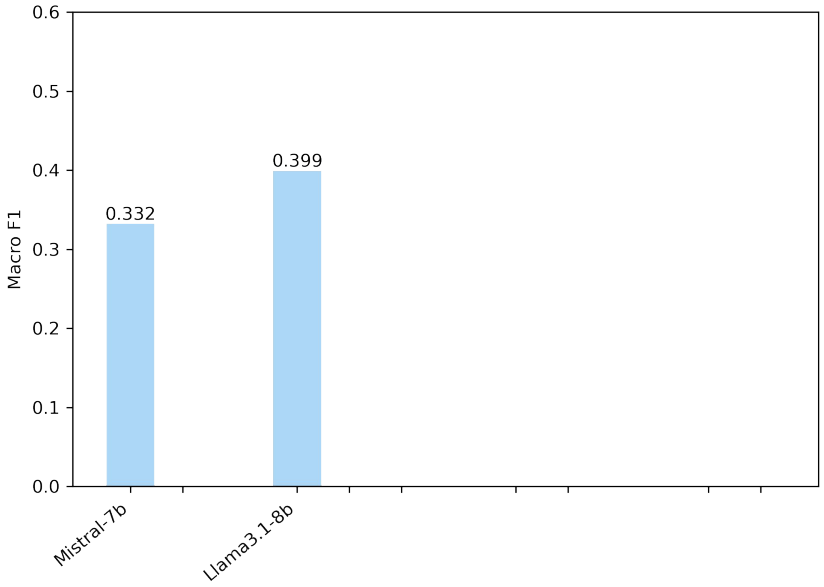
Results

Conclusion

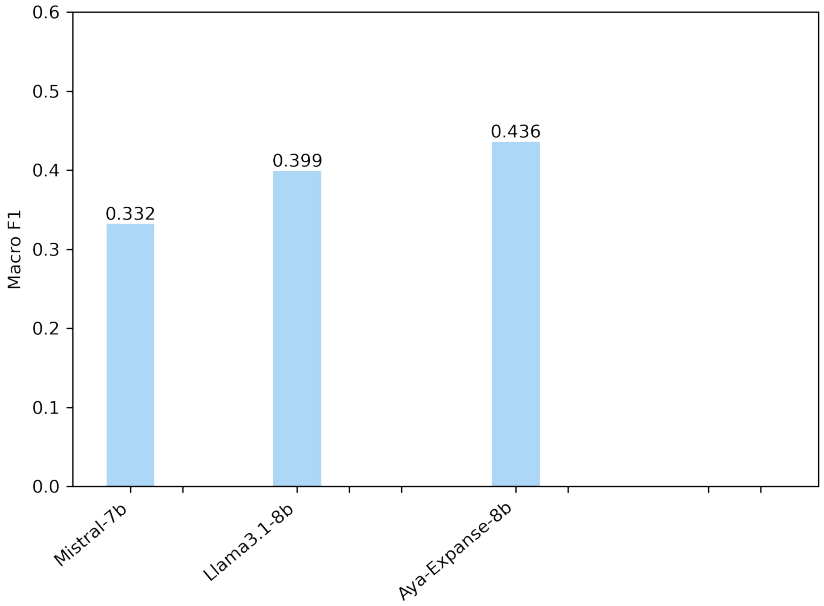
Judging Translation Candidates



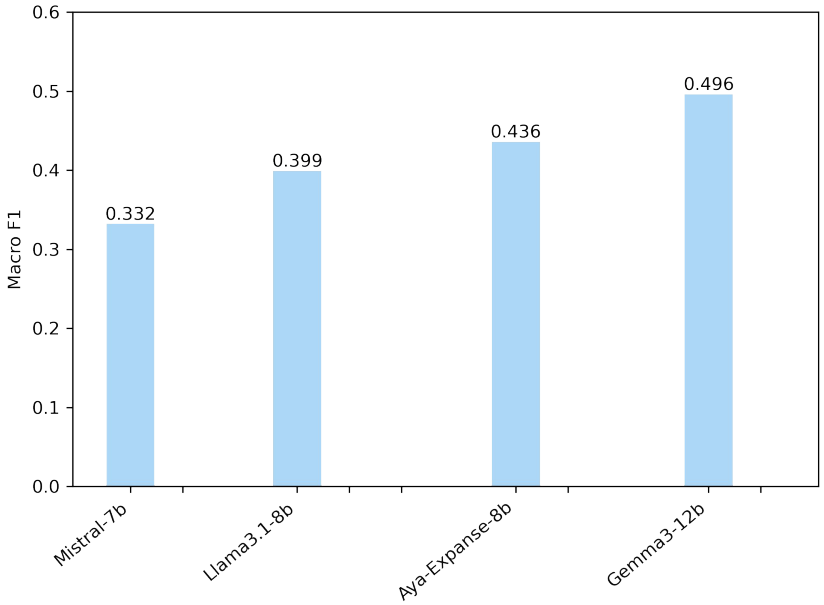
Judging Translation Candidates



Judging Translation Candidates

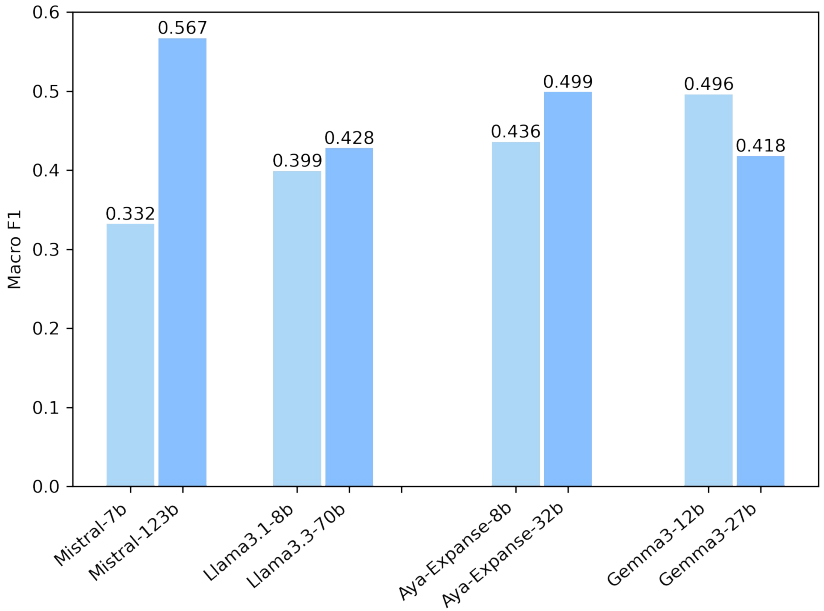


Judging Translation Candidates



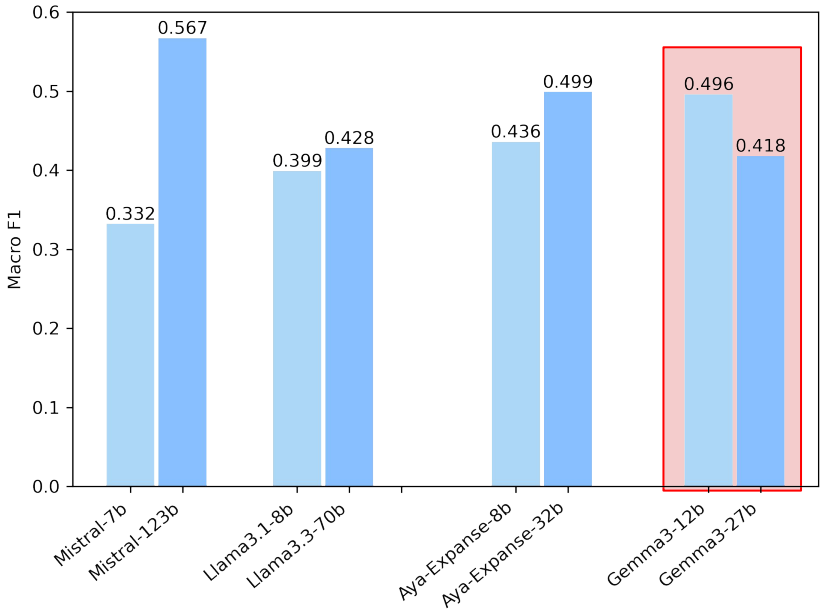
Larger and multilingual models show better results.

Judging Translation Candidates



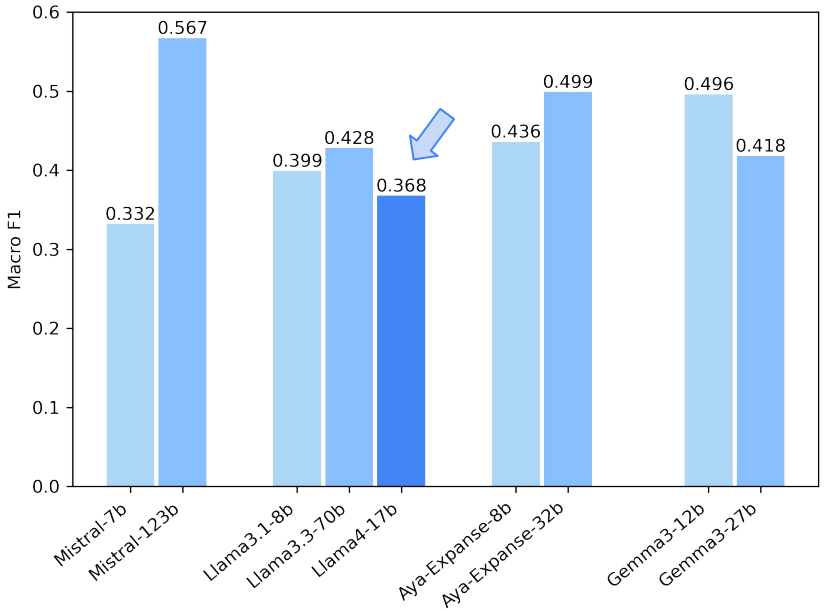
Three out of four larger model variants outperform smaller models.

Judging Translation Candidates



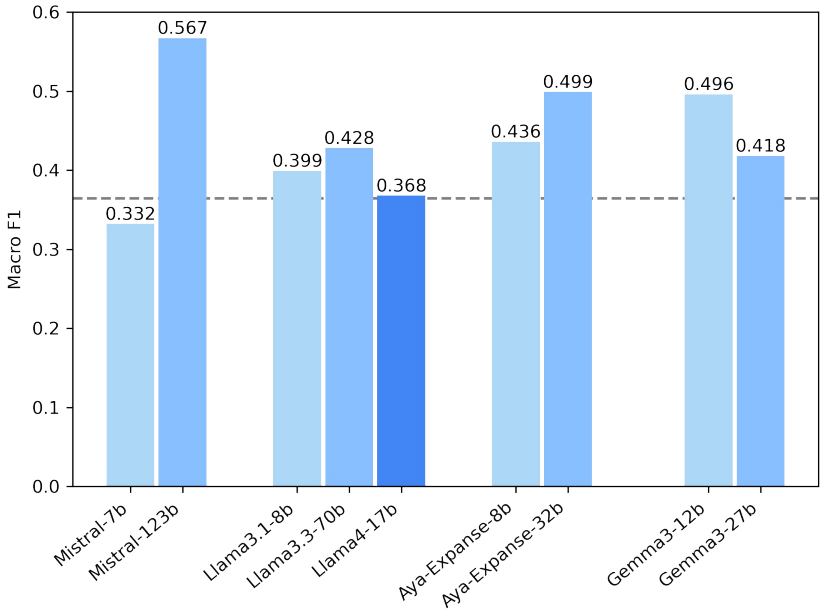
Gemma3-27b performs better at recognizing “inflections”,
but worse on other two classes.

Judging Translation Candidates



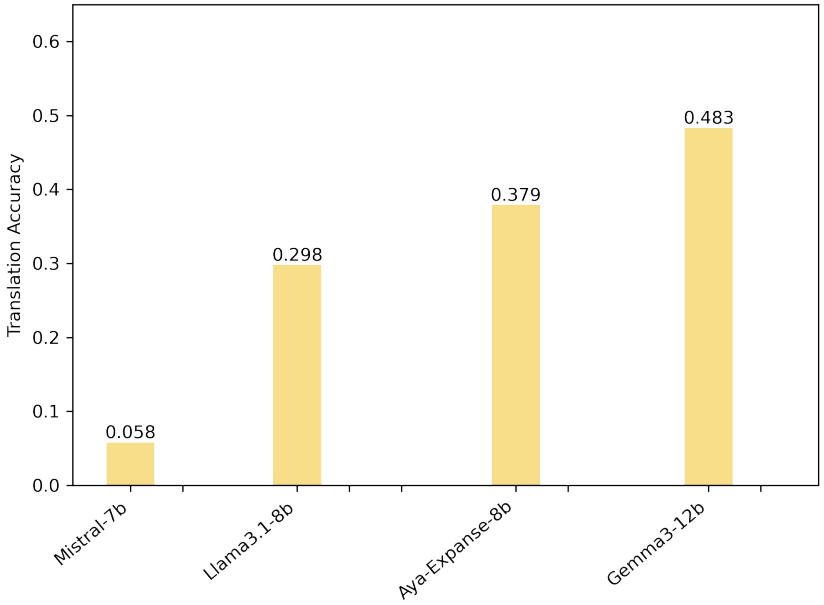
Llama4-17b performs worse than its predecessors because it often fails to follow instructions.

Judging Translation Candidates



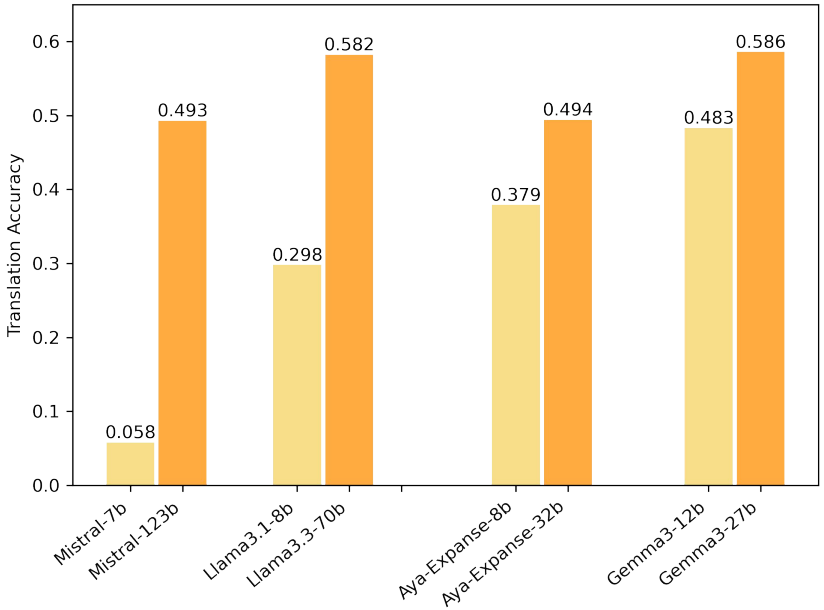
Most LLMs outperform logistic regression with string similarity features (F1: 0.364).

Dialect-to-Standard Translation



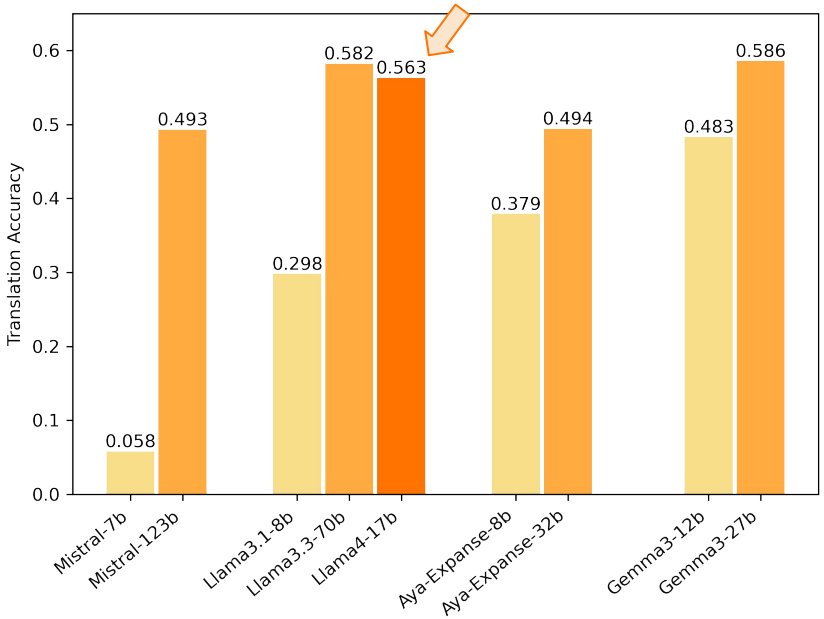
Larger models outperform smaller models (across LLM families).

Dialect-to-Standard Translation



Larger models outperform smaller models (within LLM families).

Dialect-to-Standard Translation



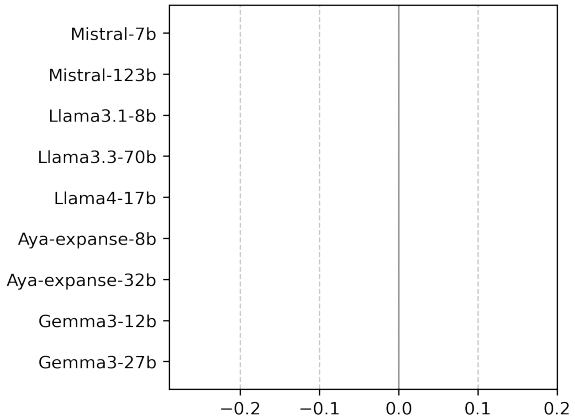
Unlike in the judgment task, Llama4-17b is able to follow instructions.

Error Analysis

Bavarian (source)	German Lemma (target)	Translation (wrong)	Error Type
Literaturwissnschoftla ("literary scholar")	Literaturwissenschaftler	Literaturwissenschaft ("literary studies")	Wrong derivative morphology.
Iabaprifung ("audit", lit. "over+test")	Überprüfung	Jahresprüfung ("yearly test", lit. "year+test")	Only part of the word is translated correctly.
Vameahrung ("proliferation")	Vermehrung	Wanderung ("hike")	Translation is entirely wrong.
…	…	…	…

(see paper for further examples)

Judging Translation Candidates



Δ = No Context - Context

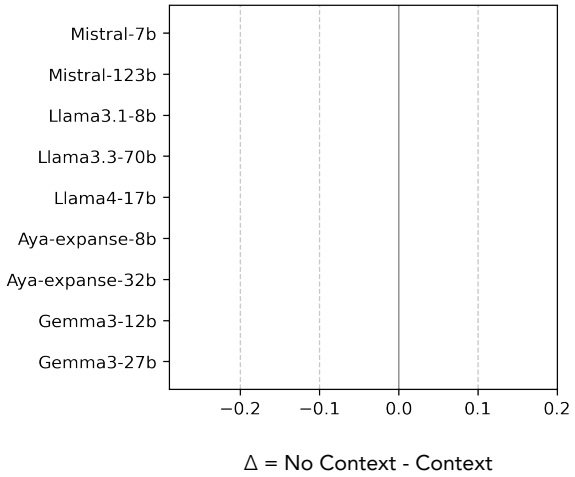
No Context

Task: Is the Bavarian term: "term_bar" a correct dialectal variant of the German term: "term_de"? Follow the given annotation guidelines.

Guidelines:
- yes: The candidate is an exact dialectal variation of the Standard German word.
- inflected: The candidate is a morphologically inflected variant of the German word.
- no: None of the two applies.

Return only "yes", "inflected", or "no".

Judging Translation Candidates



Δ = No Context - Context

No Context

Task: Is the Bavarian term: "term_bar" a correct dialectal variant of the German term: "term_de"? Follow the given annotation guidelines.

Guidelines:
- yes: The candidate is an exact dialectal variation of the Standard German word.
- inflected: The candidate is a morphologically inflected variant of the German word.
- no: None of the two applies.

Return only "yes", "inflected", or "no".

Context

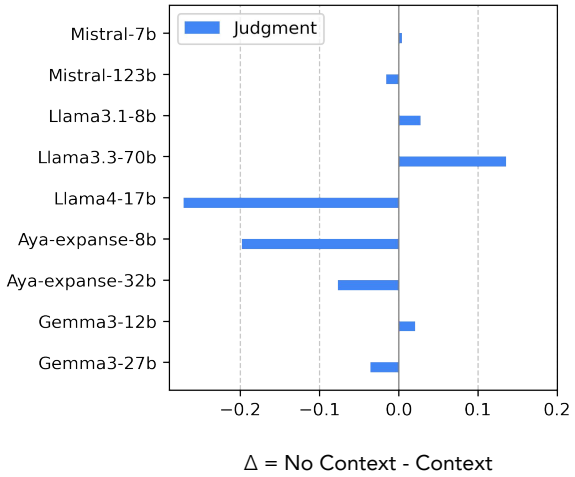
Task: Is the Bavarian term: "term_bar" a correct dialectal variant of the German term: "term_de"? Follow the given annotation guidelines.

Usage example: “####”

Guidelines:
- yes: The candidate is an exact dialectal variation of the Standard German word.
- inflected: The candidate is a morphologically inflected variant of the German word.
- no: None of the two applies.

Return only "yes", "inflected", or "no".

Judging Translation Candidates



No Context

Task: Is the Bavarian term: "term_bar" a correct dialectal variant of the German term: "term_de"? Follow the given annotation guidelines.

Guidelines:
- yes: The candidate is an exact dialectal variation of the Standard German word.
- inflected: The candidate is a morphologically inflected variant of the German word.
- no: None of the two applies.

Return only "yes", "inflected", or "no".

Context

Task: Is the Bavarian term: "term_bar" a correct dialectal variant of the German term: "term_de"? Follow the given annotation guidelines.

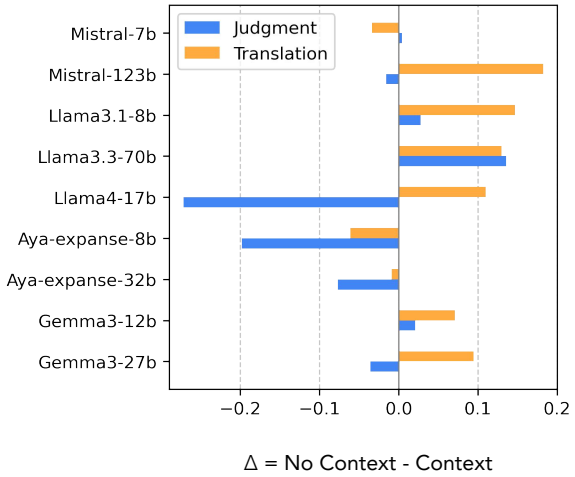
Usage example: "####"

Guidelines:
- yes: The candidate is an exact dialectal variation of the Standard German word.
- inflected: The candidate is a morphologically inflected variant of the German word.
- no: None of the two applies.

Return only "yes", "inflected", or "no".

On average, providing context deteriorates the ability to recognize dialect variants.

Dialect-to-Standard Translation



Δ = No Context - Context

No Context

Perform translation: convert the Bavarian form 'term_bar' into its Standard German equivalent. Return only the Standard German form, with no additional explanation or formatting.

No Context

Perform translation: convert the Bavarian form 'term_bar' into its Standard German equivalent. Usage example: "####". Return only the Standard German form, with no additional explanation or Formatting. Return only the Standard German form, with no additional explanation or formatting.

On average, providing context improves the ability to translate dialect terms.

Agenda

Motivation

Annotation Framework

Dialect NLP Tasks

Results

Conclusion

Conclusion

- We introduce DIALEMMA, a novel annotation framework for building dialect variation dictionaries.
- We use DIALEMMA to annotate 100K Bavarian German word pairs.
- We evaluate how well LLMs recognize and translate Bavarian words.

Paper

arxiv.org/abs/2509.17855

Code and Data

github.com/mainlp/dialemma

Thank you!