CUNI-MTIR at COVID-19 MLIA @ Eval Task 2 & 3: Multilingual Semantic Search and Machine Translation

Shadi Saleh & Hadi Abdi Khojasteh & Hashem Sellat& Pavel Pecina

Institute of Formal and Applied Linguistics
Charles University
Prague, Czech Republic

COVID-19 MLIA Eval
Jan 2021

Introduction

CUNI-MTIR team participated in the following tasks:

- Task 3: Machine Translation Task
  - In both constrained and unconstrained systems.
  - English into French, German, Spanish and Swedish.
- Task 2: Multilingual Semantic Search
  - Including its two subtasks (high precision and high recall)
  - Bilingual systems employ our MT systems from Task 3

Machine Translation Systems

Submitted systems in four pairs (EN to FR, DE, ES and SV).

- Constrained systems: using the given parallel data only.
- Unconstrained systems: using the given data plus the UFAL medical corpus

UFAL Medical Corpus

UFAL Medical Corpus is used in unconstrained systems[1], it contains parallel data from various resources:

- in-domain data:
  - EMEA corpus
  - UMLS metathesaurus
  - MuchMore corpus
  - The Corpus of Parallel Patent Applications (COPPA)
  - Titles of medical Wikipedia articles
- general-domain data:
  - CommonCrawl
  - JRC-Acquis
  - News Commentary of the Syndicate project
- Query test set and summary test set from the KConnect project

[1] http://ufal.mff.cuni.cz/ufal_medical_corpus

Constrained Systems

- MarianNMT framework to build the models
- Training on the official MLIA data only
- Data is tokenized and encoded using BPE (32K)
- The model is based on the Transformer (same parameters as Vaswani et al. 2017)
- Training is stopped after 5 iterations with no improvements

Unconstrained Systems

- Initial models are trained using the UFAL medical corpus (10 mil. sentence for each pair)
- Fine-tuned later by continuing training on the MLIA data
- Model selection: best BLEU score on the validation set

MT Results

		EN–ES	EN–DE	EN–FR	EN–SV
Constrained	BLEU	**32.9**	19.7	**34.9**	**25.1**
	ChrF	**59.1**	49.4	**60.5**	**54.1**
Unconstrained	BLEU	32.1	**20.0**	33.0	24.0
	ChrF	58.2	**49.9**	59.0	51.4

Multilingual Semantic Search

CUNI-MTIR at Task2

- Monolingual task
  - Both queries and documents in English
  - Participated in both subtasks
- Bilingual task
  - Queries in (DE, ES, FR and SV)
  - Documents only in English
  - Task reduced into monolingual by following query translation
  - Query translation using NMT systems built as in the machine translation task (unconstrained system)

Document Pre-processing and indexing

- All document fields were used while indexing (including boilerplate).
- Tokenized using Moses tokenizer.
- Covid-19 variants are mapped into one (corona virus).
- Terrier is used for indexing and querying

UDPipe: Lemmatisation

- Lemmatisation is done to reduce search space.
- Done using UDPipe (provides lemmatisation and POS tagging for 94 langauges) [2]
- Used only in Subtask2 (High-recall)

Index	#Documents	#Tokens	#Vocab
Forms	1,452,240	1,372,106,395	1,281,067
Lemmatised	1,452,240	1,364,633,452	1,244,686

[1] https://ufal.mff.cuni.cz/udpipe

Sub-tasks

- In Task-1 (High-Precision): documents are indexed in their word forms.
- In Task-2 (High-Recall): lemmatised documents are indexed instead, searching using lemmatised queries

Our Runs

Monolingual systems:

- Run1: Dirichlet model
- Run2: PL2F model (per-field normalisation)
- Run3: Query expansion using Bose-Einstein model (top 10 documents and top 5 terms)
- Run4: Query expansion using Kullback-Leiber Divergence (settings as in run3)
- Run5: Similar as run1 but conversational field used to create queries.

For the bilingual systems, translated queries are used in same runs as in monolingual systems.

Tabulka: Example of expanded queries by Bo2 and KLD Correct models.

Id	Model	Expanded Query
7	Bo2	serological tests corona virus group disease getty produce contra
	KLD	serological tests corona virus across accept dream group speech
1129	Bo2	make hand sanitizer time show currently well summit
	KLD	make hand sanitizer time class show organization currently
1135	Bo2	covid lockdown protest affect enforcement opinion sport relief
	KLD	covid lockdown protest affect scanty action violence party

Results - Subtask1

Run	English	French	German	Spanish	Swedish
Run1	68.00	**58.00**	51.33	50.00	**59.33**
Run2	46.67	36.00	34.67	50.00	32.67
Run3	48.00	38.67	39.33	40.00	40.67
Run4	42.67	37.33	35.33	38.67	34.67
Run5	**68.67**	54.00	**52.00**	**54.00**	57.33

Results of Subtask1 (High-Precision) in terms of P@5 in percentages

Results - Subtask1

Run	English	French	German	Spanish	Swedish
Run1	**60.21**	**53.17**	45.01	45.93	51.50
Run2	40.63	30.94	31.68	28.88	29.26
Run3	37.14	31.76	31.64	34.35	31.86
Run4	33.05	30.35	26.66	33.34	27.58
Run5	58.34	49.64	**48.40**	**51.93**	**53.67**

Results of Subtask1 (High-Precision) in terms of NDCG@5 in percentages

Results: Subtask2

Run	English	French	German	Spanish	Swedish
Run1	59.33	50.67	46.00	50.67	42.67
Run2	46.60	36.00	34.67	34.00	32.67
Run3	36.00	26.67	24.67	20.67	24.00
Run4	35.33	28.00	26.67	25.33	24.67
Run5	**69.33**	**55.33**	**56.67**	**62.00**	**56.67**

Results of Subtask2 (High-Recall) in terms of P@5 in percentages

Results: Subtask2

Run	English	French	German	Spanish	Swedish
Run1	48.48	40.35	35.76	41.45	36.32
Run2	40.63	30.94	31.68	28.88	29.26
Run3	30.20	20.40	19.56	15.26	18.82
Run4	28.55	22.00	19.51	19.08	18.09
Run5	**57.02**	**45.32**	**48.41**	**50.64**	**45.59**

Results of Subtask2 (High-Recall) in terms of NDCG@5 in percentages

Results: Comparison

subtask	English	French	German	Spanish	Swedish
Subtask1	**62.23**	**26.48**	19.24	30.84	**29.84**
Subtask2	58.51	26.08	**22.50**	**33.79**	29.54

Comparison between the two subtasks in terms of Precision Averages at 0.10 recall (Run5 only)

Future Work

- For Machine Translation:
  - Investigate advanced corpus filtering methods for data cleaning.
  - Employ Back-translation and transfer learning.
- For Multilingual-Search:
  - Employ embedding-based similarity function for retrieval.