MSc Language Technology

Thesis Defense

(10/1/2025)

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

IEA ILSP

Keypoint Analysis in Greek
A new dataset and baselines

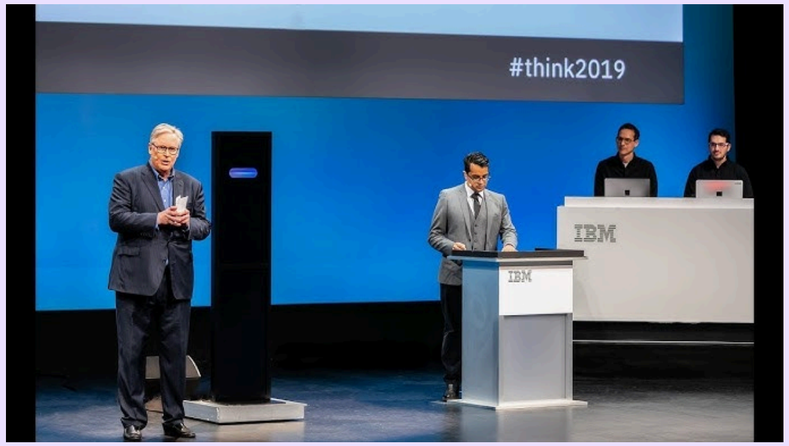
Student Name: Kleopatra Karapanagiotou
Supervisor: Dimitris Galanis, Researcher C' (ILSP)

Introduction

Keypoint Analysis (KPA) is a novel summarization framework (Haim et al.,2020a,b)

Aim

- Given a collection of opinionated texts
- Select and summarize the most significant points and rank them by their prevalence

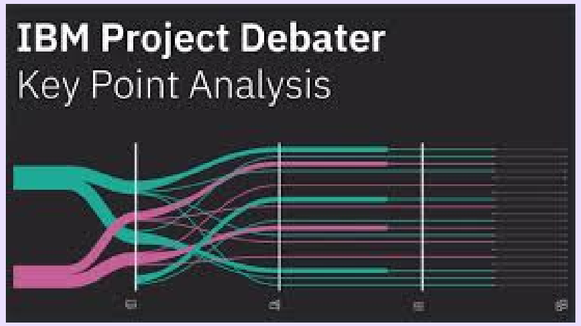


Inspiration

IBM Project Debater

Introduction

KPA: Both Textual and Quantitative Summary



There is currently too much wealth inequality in the world.	153

- → With gaps in wealth distribution rapidly growing, I think it's time to fix the problem.
- → The current state-of-affairs is very bad. Too few people hold too much of the world wealth.
- → Wealth has concentrated in the hands of a select few while the rest of the people in the world are poor. So we should redistribute wealth.

Key Point Analysis - Top Key Points

Topic: *It is time to redistribute wealth*

Key Points - For	#Args	Key Points - Against	#Args
There is currently too much wealth inequality in the world.	153	Redistributing wealth would discourage some people from working hard	107
Redistribution of wealth will allow those with less opportunity to achieve success	75	Wealth should be earned not taken from others.	54
The possibility of accumulating wealth allows economic growth and innovation	43	redistributing wealth would harm the investment into innovation	50
Inequality hurts economic growth, especially high inequality in rich nations.	36	Forced redistribution of wealth is a form of theft.	29
Great inequities in wealth leaves the poorest at greatest risk	32	Wealth is limited and redistribution is not a sustainable solution.	24

KPA Applications/ Use cases

In a time of great polarization KPA can benefit

- citizens and government
- employees and managers
- businesses and customers

Towards

- better communication
- informed/data-driven decision making

KPA Shared Task 2021 (IBM)

ArgKP-2021 dataset

- domain: argumentation
- 31 debatable topics
- 2 stances (pro/con)
- ~7000 high quality arguments
- ~280 keypoints (by domain expert)
- Annotation of matching/non matching <arg,kp> pairs (crowdsourcing)

	Train	Validation	Test
Num topics	24	4	3
Num. arguments	5583	932	723
Num. key points	207	36	33
num <arg.,kp.,label> pairs	20635	3458	3426
num <arg., kp.> pairs (+undecided pairs)	24454	4211	3923
num matching pairs	17 %	18%	14 %
num non-matching pairs	67 %	64 %	73 %
num undecided pairs	14%	18%	13%

KPA Subtasks

Keypoint Matching (KPM)

Keypoint Generation (KPG)

KPM

Description

In each topic-stance combination:

- Report for each argument its match score for each of the reference key points

Evaluation Metrics

mean Average Precision (mAP)

In each topic-stance combination:

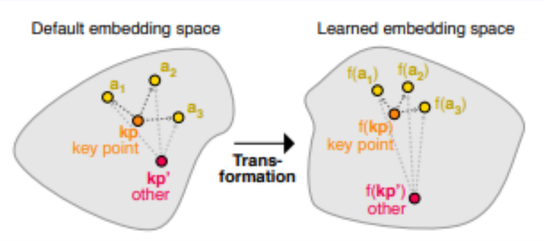
1.Match each argument with highest scoring kp.
2. Keep 50% of the top-ranked pairs
3. Calculate Average Precision for remaining pairs based on labelled data
4. Obtain mAP (macro-average over all topic-stance combinations)
5. mAP(strict): undecided pairs labelled as “no-match”
6. mAP(relaxed): undecided pairs labelled as “match”
7. Final rank: **avgmAP** (of mAP(s) and mAP(r))

KPM Solutions

During Shared Task (NLU)

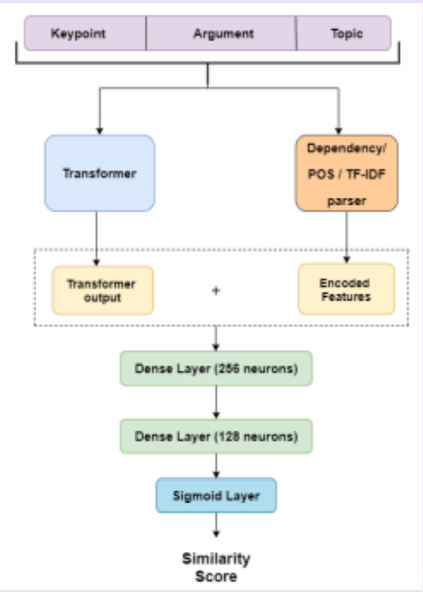
Sentence similarity

- 1st ranked: SMatchToPR
- Trained SBERT model
- Input : <Arg, kp+topic>
- Encoding: RoBERTa-large
- Train in Siamese NN
- loss: contrastive
- avgmAP : 0.858



Sentence pair Classification

- 3rd ranked: Enigma
- classifier architecture
- Input: concatenated <arg,kp,topic>
- Embedded with DeBERTa-large
- Output concatenated with their POS features
- Fed to two more dense layers
- Output: similarity/matching score
- avgmAP: 0.833



Post Shared Task (NLG)

Text classification in a generative manner

1. **F**lan-T5-XL (Finetuning with QLoRa, 1 epoch)
- Prompted to reply with yes/no answers
- Results comparable to a DeBERTa-large

2. GPT 3.5,4 (0-shot prompting)
- avgmAP: 0.17

Prompt 3: ChatGPT open book, KPM prompt

For the claim of **{claim}**, indicate for each of the following argument/key point pairs whether the argument matches the key point. Return a JSON object with just a "match" boolean per argument/key point pair.
ID: **{pair id}** Argument: **{argument}** Key point: **{key point}** (*up to* B_{KPM} *times*) . . .

KPG

Description

For each set of arguments supporting or contesting a given topic:

- Generate a set of keypoints

Evaluation

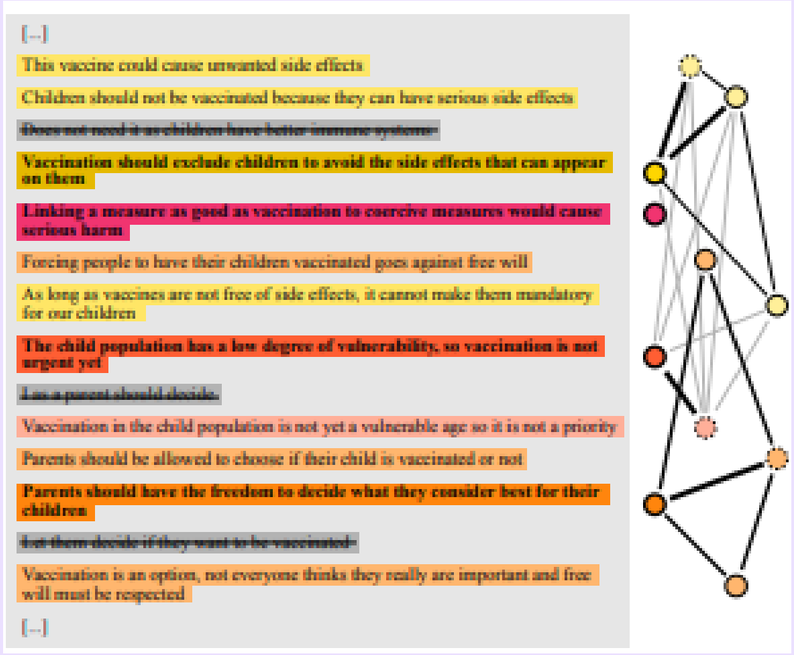
- human (keypoint quality)
  1. clarity of stance
  2.coverage
  3. redundancy

- automatic:
  1. n-gram based (ROUGE)
  2. embedding-based(BERTScore)
  3. coverage
  4.redundancy

KPG Solutions

During Shared Task

- Mostly extractive
- 1st ranked: SMatchToPR
- Graph-based (PageRank)
- nodes =high-quality arguments
- edges=match scores betw. arg pairs
- arguments with the most connections are chosen as keypoints



Post-Shared Task

Clustering-based

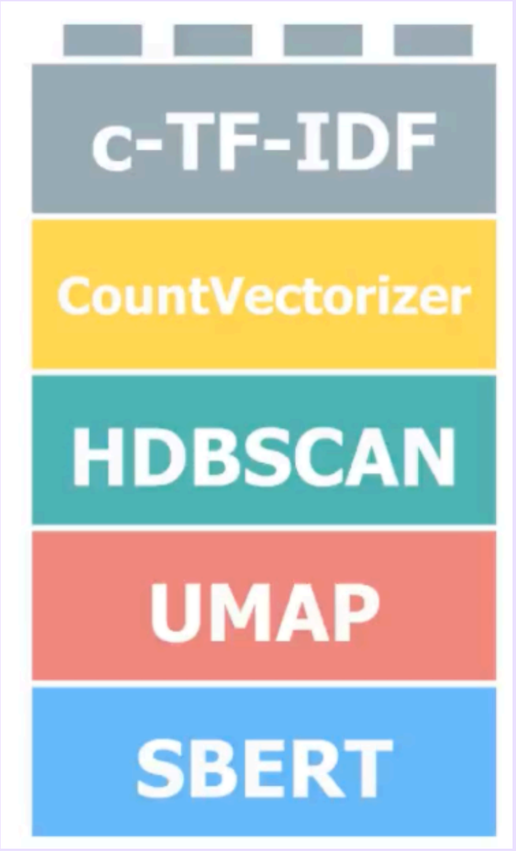
Extractive

Abstractive

Abstractive KPG

Argument clustering

BERTopic



all docs in a cluster-1 single doc tf-idf to find the most important words for each cluster

Tokenize (BoW matrix)

Cluster reduced embeddings

Reduce dimensionality

Embed docs (Transformers)

Generative LM

- Finetune encoder-decoder models (Flan-T5, Pegasus-XSUM) for abstractive summarization
- Generate a keypoint for each cluster

Thesis objectives

In view of

- The recent advancements in Greek NLP:
  (GreekBERT (2020), Greek T5 series(2024) , Meltemi-7b(2024))

Our aim is to

- Explore KPA in a non-English setting (first attempt in KPA)
- Develop the Greek version of KPA
- Provide appropriate datasets
- Set baseline solutions for both KPA subtasks

Challenge

ArgKP-2021 Dataset translation

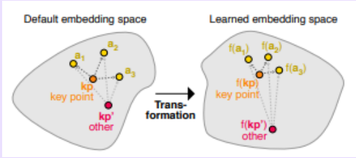
Set	num. pairs	Translation
Train	24454	0-SHOT MT (MADLAD400-3B-mt) • T5-based multilingual MT model • Trained on 1T tokens (covers 450 languages) • For Greek 3B competitive to 7.2B,10.7B models
Dev	4211	human
Test	3923	human

KPM preliminary experiments

- **Re-implement 2 existing KPM solutions with available Greek models**

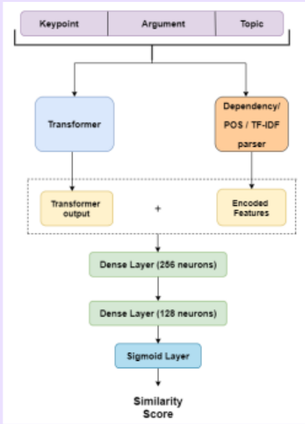
SMatchToPR (SBERT model)

- RoBERTa-large
- 10 epochs
- trained on labelled ArgKP 2021 train set (20.635 pairs)



Enigma(custom classifier)

- DeBERTa-large
- 3 epochs
- no POS features
- trained on labelled ArgKP 2021 train+val set (24.093)



Greek monolingual edition of Google's BERT PLM

- base version (110m params)
- 29 GB Greek data
- Same training setup

ArgKP-2021 (test set)

Model	Measures					
	EN (BERT)			GR (GreekBERT)		
	mAP strict	mAP relaxed	Avg mAP	mAP strict	mAP relaxed	Avg mAP
Enigma replication	79.92	91.70	85.81	78.96	85.07	82.01
SMatchtoPR-replication	81.83	88.16	85.00	80.15	90.10	85.12

Proposed KPM Baseline

Inspired by recent work (Benayas et al., 2024)

- Finetuning Decoder-only models (LLaMa-2 7B) gives competent results in NLU tasks (intent classification/sentiment analysis) to encoder-only (RoBERTa)

And recent advancements in Greek NLP:

Meltemi (ilsp)

- First decoder-only model for Greek
- Built on top of Mistral-7b
- Developed as bilingual model with various techniques

Proposed KPM Baseline

Decoder-based finetuned classifier

- Model : Meltemi-7B-v1 (foundation)
- 4-bit Quantization (bitsAndbytes)
- PEFT (LoRa): freeze weights, add light low-rank matrices for training)
- Train data: labelled ArgKP-2021 train set (20635 pairs)
- Match score= the probability of the positive class (class_1)
- class weights

Hardware used (Kaggle resources)

- Number of nodes: 1
- Number of GPUs per node: 1
- GPU type: NVIDIA P100
- GPU memory: 16GB

classes: 2	optimizer: paged Adam optimizer
epochs: 1	Seed: 42
max_seq_length: 512	LoRA r :8
batch_size: 16	LoRA alpha: 8
Gradient Accumulation Steps: 2	LoRA dropout: 0.0
learning_rate: 1e-4	LoRA bias: ‘none’
lr_scheduler_type: linear	target_modules: q_proj, v_proj
Weight Decay: 0.01	task_type: "SEQ_CLS"
M. G. Norm: 0.3	Loss: Binary Cross Entropy
trainable parameters: 3,416,064 (~5% of the original model)	
training duration (+/- 20 hours)	

Experiments

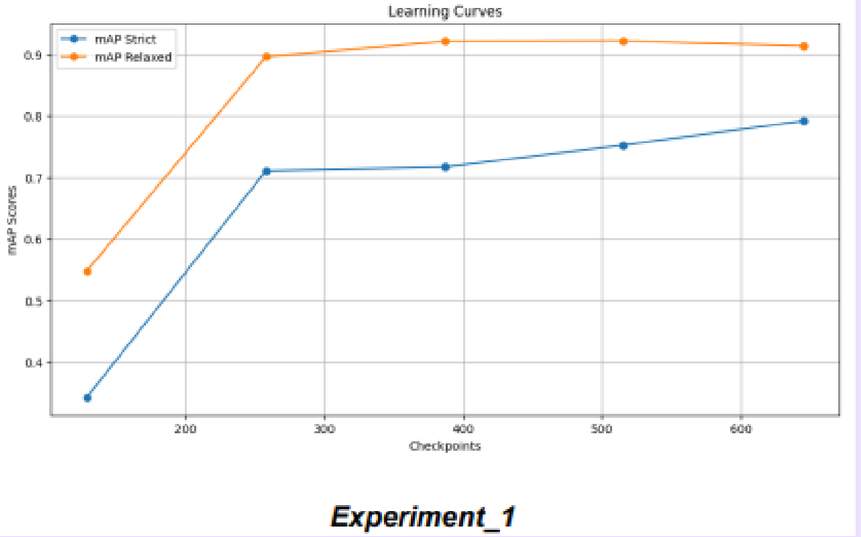
Experiment_1

Input : Keypoint: <kp>; Argument: <arg>
mAP(s): 79.18
mAP(r): 91.37
avg mAP: 85.27

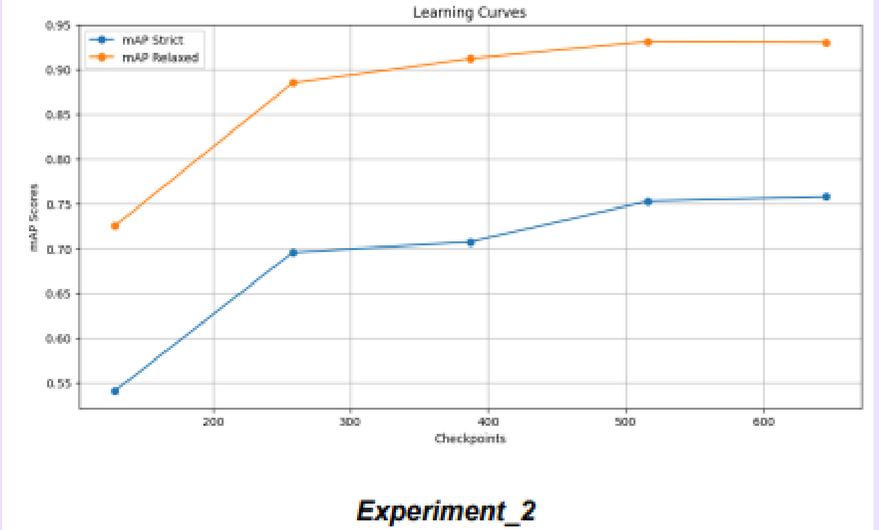
Experiment_2

Input : Keypoint: <kp>; Argument: <arg>; Topic:<topic>
mAP(s): 75.74
mAP(r): 93.00
avg mAP: 84.37

ArgKP-2021 dev set



Experiment_1



Experiment_2

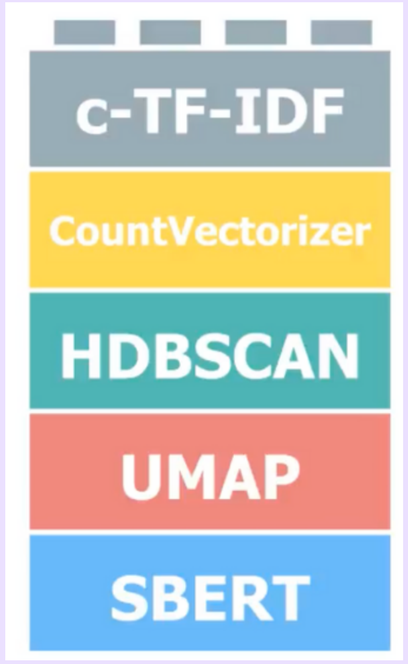
KPM Results on Test set

- Dev set results validated
- SMatchToPR outperforms Enigma
- The “topic” context does not help Meltemi-classifier in the KPM task
- Enigma GreekBERT-base classifier (avg mAP: 82.01)
- Meltemi-base-classifier (avg mAP: 87.73)

Experiments	Num. training instances	mAP (strict)	mAP (relaxed)	avg mAP
SmatchtoPR (arg,topic+kp)-GreekBERT	20.635	80.15	90.10	85.12
Enigma (kp,arg,topic)-GreekBERT	24.093	78.96	85.07	82.01
Meltemi-base (kp, arg) - (weighted)	**20.635**	**83.86**	**94.27**	**89.06**
Meltemi-base (kp,arg,topic) - (weighted)	20.635	81.78	93.68	87.73

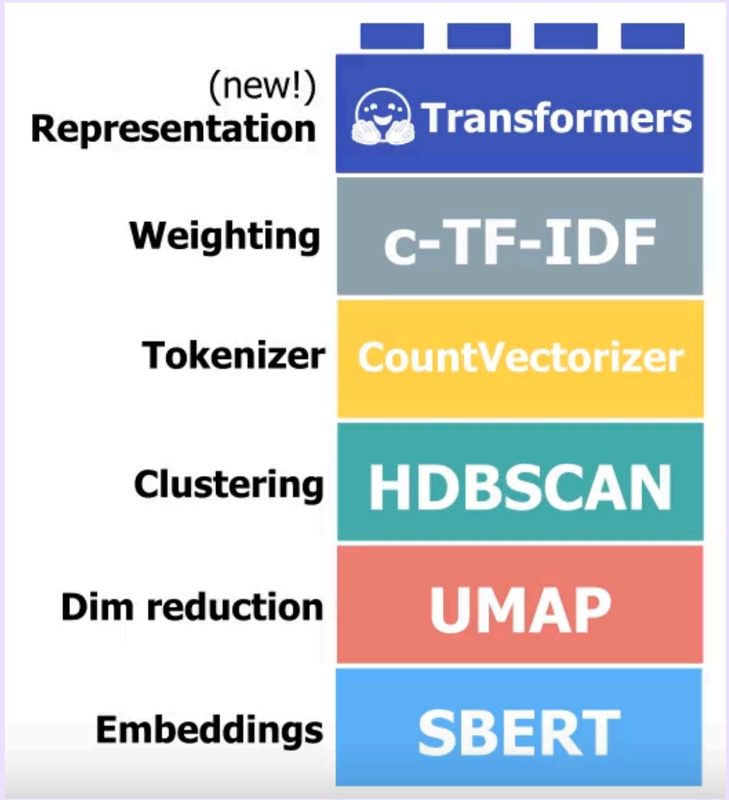
KPG METHODS

Previous Work



+

Ours



Main idea

"Instead of keywords, generate respresentative sentences for each cluster"

KPG EXPERIMENTS

Representation model finetuning

GreekWiki

- GreekT5 series model (encoder-decoder)
- umt5-base (580M params)
- Finetuned on Greek Wikipedia articles (encyclopedic article summarization)

Meltemi-base

- decoder-only
- foundation model
- 7B params
- v1.5

Meltemi-Instruct

- decoder-only
- instruction-tuned model
- 7B params
- v1.5

KPG Metrics

ROUGE (1,2,L)

For each topic/stance combination:

- compute R1,2,L of each predicted with each reference keypoint
- take the average scores for each metric

Take the average of all topic-stance combinations for each metric
Stemming: Greek Stemmer

BERTScore

For each topic/stance combination:

- compute Precision,Recall, F1 of each predicted with each reference keypoint
- take the average scores for each metric

Take the average of all topic-stance combinations for each metric
Default model for non-English languages (mBERT)

Deduplication

- keypoints ranked in descending order
- Drop duplicate keypoints with similarity score > 0.9 with a higher-ranked keypoint

0-shot experiments

Start with GreekWiki Model-specific
Prompt:
“summarize: ” : <input text>
Decoding strategy:
Greedy/ Beam_3

	GreekWiki			Meltemi base (v1.5)			Meltemi Instruct (v1.5)		
0-shot summarization-specific prompt (greedy)	**Rouge** 1: 14.7 2: 5.3 L: 13.9	**Bertscore** P: 67.1 R : 70.7 F1: 68.8	**Avg token count** 20.1	**Rouge** 1: 15.4 2: 5.9 L: 14.6	**Bertscore** P: 66.3 R: 72.5 F1: 69.2	**Avg token count** 19.41	**Rouge** 1: 15.4 2: 6.0 L: 14.6	**Bertscore** P: 66.9 R: 73.6 F1: 70.0	**Avg token count** 24.16
0-shot summarization-specific prompt (beam_3)	1: 12.8 2: 4.8 L: 12.5	P: 66.5 R: 70.1 F1: 68.2	19.72	1: 15.0 2: 5.0 L: 14.3	P: 66.9 R: 72.7 F1: 69.6	18.61	1: 15.8 2: 5.3 L: 14.8	P: 66.3 R: 73.6 F1: 69.7	24.45

- Meltemi outputs (semantically more coherent compared to GreekWiki)
- Greedy outputs (semantically incoherent for GreekWiki)
- Continue with beam_3
- Optimize Meltemi-instruct

Prompt Engineering Experiments

	Custom Prompts (GR)
Prompt_1	Γράψε μια σύντομη πρόταση ως περίληψη για το παρακάτω κείμενο: [ARGUMENTS] Περίληψη:
Prompt_2	Τα παρακάτω επιχειρήματα υποστηρίζουν ή αντικρούουν το θέμα. Συμπλήρωσε ένα συνοπτικό keypoint, καταγράφοντας την κεντρική ιδέα των επιχειρημάτων σε μία πρόταση. [ARGUMENTS]
Prompt_3	**Παρακάτω θα δεις μερικά επιχειρήματα υπέρ ή κατά για ένα συγκεκριμένο θέμα: [ARGUMENTS] Με βάση τα παραπάνω, γράψε μια σύντομη πρόταση που να συνοψίζει αυτά τα επιχειρήματα σε ένα keypoint, ακολουθώντας το μοτίβο: θέμα: <keypoint>**

Meltemi-Instruct(beam_3)- ArgKP-2021 Dev Set

Custom Prompts	**Automatic evaluation measures**						
	ROUGE			BERTscore			Avg token count
	1	2	L	P	R	F1	
Prompt_1	14.7	4.9	13.7	65.6	73.4	69.3	26.9
Prompt_2	14.8	5.0	13.8	65.8	73.6	69.4	30.0
Prompt_3	**16.9**	**6.0**	**15.9**	**67.6**	**73.5**	**70.4**	**20.96**

- Prompt_1+2: Complex outputs
- Prompt_3: shorter, more keypoint-like outputs

Final settings

- beam_3 decoding strategy
- GreekWiki: model-spefic prompt
- Meltemi-base & -Instruct: Prompt_3

Few-shot experiments

- Best performing model in 0-shot: Meltemi-Instruct (beam_3,Prompt_3)
- 4-8-16 demonstrations (from the ArgKP-2021 (mt) train set)
- Each demonstration shows 4 arguments of each keypoint
- For each topic, show 1 demonstration for each stance (positive/negative)

topic (urbanization) stance (pro)

```
<|system|>
Είσαι το Μελτέμι, ένα γλωσσικό μοντέλο για την ελληνική γλώσσα. Είσαι ιδιαίτερα βοηθητικό προς την χρήστρια ή τον χρήστη και δίνεις σύντ
ομες αλλά επαρκώς περιεκτικές απαντήσεις. Απάντα με προσοχή, ευγένεια, αμεροληψία, ειλικρίνεια και σεβασμό προς την χρήστρια ή τον χρήστ
η.</s>
<|user|>
Παρακάτω θα δεις μερικά επιχειρήματα υπέρ ή κατά για ένα συγκεκριμένο θέμα:
- θα μπορούσε να προσφέρει περισσότερες θέσεις εργασίας και να προωθήσει την ανάπτυξη των επιχειρήσεων.
- Η αστικοποίηση επιτρέπει στη χώρα να εξοικονομήσει χρήματα, επειδή μπορούν να χτίσουν κεντρικές υπηρεσίες όπως νοσοκομεία και σχολεία.
- η αστικοποίηση φέρνει θέσεις εργασίας και είναι καλή για την οικονομία.
- η αστικοποίηση μπορεί να φέρει τα πολύ απαραίτητα χρήματα σε μια πόλη.
Με βάση τα παραπάνω, γράψε μια σύντομη πρόταση που να συνοψίζει αυτά τα επιχειρήματα σε ένα keypoint, ακολουθώντας το μοτίβο:
θέμα: <keypoint></s>
<|assistant|>
θέμα:Η αστικοποίηση ωφελεί την οικονομία</s>
<|user|>
Παρακάτω θα δεις μερικά επιχειρήματα υπέρ ή κατά για ένα συγκεκριμένο θέμα:
[DOCUMENTS]
Με βάση τα παραπάνω, γράψε μια σύντομη πρόταση που να συνοψίζει αυτά τα επιχειρήματα σε ένα keypoint, ακολουθώντας το μοτίβο:
θέμα: <keypoint></s>
<|assistant|>
```

Dev results

- More shots, better scores, more keypoint-like outputs
- 4-shot competitive to 8- and 16-shot
- Due to computational restrictions, we continue with 4-shots

Meltemi Instruct (1.5)	Rouge			BERTScore			Avg token count
	1	2	L	P	R	F1	
0-shot Prompt_3 (beam_3)	16.9	6.0	15.9	67.6	73.5	70.4	20.96
4-shot Prompt_3 (beam_3)	23.5	10.1	22.6	72.8	75.7	74.2	10.68
8-shot Prompt_3 (beam_3)	24.8	11.0	24.3	74.7	75.4	75.0	8.23
16-shot Prompt_3 (beam_3)	**25.3**	**11.9**	**24.9**	**75.1**	**75.1**	**75.0**	**7.92**

KPG Results on Test set

1. GreekWiki incompetent for KPG in 0-shot setting. (definition-like, semantically unmeaningful outputs)
2. Meltemi-base repeats arguments from input prompt
3. Meltemi-Instruct 0-shot produces syntactic paraphrases of arguments in the imput prompt
4. Meltemi-Instruct 4-shot significantly outperforms all other settings (indications of hallucinations, topic-like keypoints)

Experiment setting	**Rouge**			**Bertscore**			**Avg token count**
	1	**2**	**L**	**P**	**R**	**F1**	
GreekWiki (0-shot)	12.3	3.6	11.0	66.0	67.5	66.7	24.08
Meltemi base (0-shot)	13.2	2.3	11.5	66.9	69.1	68.0	19.5
Meltemi Instruct (0-shot)	15.8	4.6	14.1	68.0	70.6	69.2	20.5
Meltemi Instruct (4-shot)	**20.2**	**8.0**	**19.1**	**74.0**	**72.8**	**73.4**	**10.89**

Best KPG model example output

Topic: Οι ΗΠΑ είναι μια καλή χώρα για να ζει κανείς/ The USA is a good country to live in
Stance: -1

Language	GR	EN
Predicted	1. Οι ΗΠΑ αντιμετωπίζουν σοβαρά κοινωνικά και οικονομικά προβλήματα 2. Οι Ηνωμένες Πολιτείες έχουν ένα ιστορικό φυλετικών διακρίσεων 3. Διακρίσεις στην αγορά εργασίας 4. Η εγκληματικότητα στις ΗΠΑ είναι ένα σημαντικό πρόβλημα 5.Η υγειονομική περίθαλψη και η εκπαίδευση είναι πολύ ακριβές στις ΗΠΑ 6. Το σύστημα υγείας των ΗΠΑ είναι πολύ ακριβό και δημιουργεί διακρίσεις	1. The US faces serious social and financial problems 2. The United States has a history of racial discrimination 3. Discrimination in the labor market 4. Crime in the United States is a significant problem 5.Healthcare and education are very expensive in the U.S. 6. The US healthcare system is very expensive and discriminatory
References	1. Οι Ηνωμένες Πολιτείες έχουν άδικες πολιτικές στους τομείς της υγείας και της εκπαίδευσης 2. Οι Ηνωμένες Πολιτείες έχουν ένα προβληματικό/διχαστικό πολιτικό σύστημα 3. Οι Ηνωμένες Πολιτείες έχουν υψηλή φορολογία και υψηλό κόστος διαβίωσης 4. Στις Ηνωμένες Πολιτείες υπάρχει ξενοφοβία και ρατσισμός 5. Οι Ηνωμένες Πολιτείες έχουν ανισότητες και φτώχεια 6. Στις Ηνωμένες Πολιτείες δεν υπάρχει ασφάλεια 7. Στις Ηνωμένες Πολιτείες υπάρχει η αρνητική κουλτούρα	1. The US has unfair health and education policies 2. The US has a problematic/divisive political system 3. The US has high taxation/high costs of living 4. The US is xenophobic/racist 5. The US has inequality/poverty 6. The US is unsafe 7. The US has a negative culture

Achievements

1. **ArgKP-2021-GR dataset**
2. KPM baselines (3) for Greek (Greek Enigma, Greek SMatchToPR, **Meltemi-KPM-classifier**)
3. Abstractive KPG baseline (1) for Greek (**Clustering-based Meltemi-Instruct 4-shot**)

Both Meltemi-based KPM, KPG baselines gave promising results

Future Directions

KPM

1. More PEFT methods
2. Classification as NLG task
3. KPM the “bottleneck” of KPA in real-life applications, explore more efficient solutions

KPG

1. Improve Argument clustering methods : Iterative clustering
2. Improve embedding quality (with a Greek Sentence-transformer model, not yet available)
3. Meltemi-finetuning
4. Establish a more clear evaluation framework

References

- Benayas, A., Sicilia, M. A., & Mora-Cantallops, M. . A Comparative Analysis of Encoder Only and Decoder Only Models in Intent Classification and Sentiment Analysis: Navigating the Trade-offs in Model Size and Performance. Language Resources and Evaluation. 2024. https://doi.org/10.1007/s10579-024-09796-y

- Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From Arguments to Key Points: Towards Automatic Argument Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4029–4039, Online. Association for Computational Linguistics.

- Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative Argument Summarization and Beyond: Crossdomain Key point Analysis. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 39–49, Online. Association for Computational Linguistics. 2020.

Thank you

QnA