Finding Good Answers in Online Forums: Community Question Answering for Bulgarian

Tsvetomila Mihaylova, **Ivan Koychev**
FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
Preslav Nakov
ALT Research Group, Qatar Computing Research Institute, HBKU, Doha, Qatar
Ivelina Nikolova
IICT, Bulgarian Academy of Sciences, Sofia, Bulgaria

Content

What is CQA

Task Formulation

Method

Data

Experiments

Conclusion

Content

What is CQA

Task Formulation

Method

Data

Experiments

Conclusion

Think of the last time...

... you had to search for information in a forum...

Did you find it immediately?

Did you have to read a lot of useless content?

Purpose of CQA

Finding useful responses to questions from user-generated content.

User-Generated Content

Lots of Information - some of it is useful

Lots of noise

Grammar, speech, expression…

Need to extract meaningful info from it

Example for CQA

On next slides (*in Bulgarian - translation follows*)...

(From BG-Mamma - largest online forum in Bulgaria)

Question subject

Въпроси относно камина Ерато

💙 Харесва ми 0   f Share 0

👁 970 прегледа   💬 13 отговора

💬 Отговор   Включете известията   Още ▾

0 Потребители и 1 Гост преглежда(т) тази тема.

Question user

Майсторката ▾   Ентусиаст

Въпроси относно камина Ерато

♥0 пн, 28 сеп 2015, 19:34

Цитирай

Враца

Мнения: 112

Моля тези от вас, които имат такава камина да се включат с отговор на няколко въпроса:

1. Запалихте ли вече камините. Първоначално само вечер ли?

2. Какви настройки сте направили?

3. Имате ли някакво ръководство?

4. Какви пелети ползвате?

Предварително благодаря на всички 😄

Question text

Цитирай

8

змия ▼   Ентусиаст   ♥ 0 пн, 28 сеп 2015, 21:27 **#1**

Re: Въпроси относно камина Ерато   Цитирай

София

Мнения: 3 960

Имам Пони9 на Ерато.

Днес я запалих за 2 часа.

Пелетите са български от Разлог, но имаме още 2-3 торби от тях. За тази зима сме поръчали етрополски пелети 2,5 тона.

Засега не сме настройвали нищо. Миналата година бяхме настроили да се включва сутрин в 5:30, после по някое време се изключваше, пак се включваше и т.н., но не помня подробности.

Имам книжка с инструкции.

Comment text

Цитирай

Майсторката ▼   Ентусиаст   ♥ 0 пн, 28 сеп 2015, 22:33 **#2**

Re: Въпроси относно камина Ерато   Цитирай

Благодаря за отговора. Използвали ли сте някакъв екорецим?

Враца

Мнения: 112

Цитирай

9

10

Position of the comment in the thread

ЗМИЯ ▼ Ентусиаст   ♥ 0 вт, 29 сеп 2015, 10:1. #5

Re: Въпроси относно камина Ерато   Цитирай

София

Мнения: 3 960

Първата зима спираше, но после от фирмата, откъдето я купихме, й промениха настройките и сега не спира.

Проблемът със самоизключването бе, че трябва температурата да падне с 2 градуса под зададената, за да се включи.

По този начин се получаваха големи температурни амплитуди.

Цитирай

Майсторката ▼ Ентусиаст   ♥ 0 вт, 29 сеп 2015, 10:20 #6

Re: Въпроси относно камина Ерато   Цитирай

Враца

Мнения: 112

Мн полезно инфо:-P а на какъв режим сте сега?

Цитирай

11

Example for CQA

And now, translated to English…

Prepared for processing…

And annotated…

Question Subject	Question Text
Questions about fireplace Erato	Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol:

Comments

Position	Relevance	Comment Text
1	Good	I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions.
2	Bad	Thanks for the reply. Have you used any ekorecim?
3	Bad	I have no idea what that is. :shock:
4	Bad	Buhahaha :D These jokes become, when you write in the phone. I meant the ECO MODE : hug: by the way your fireplace when it reaches a certain temp stops work?
5	Good	The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes.

Question Subject	Question Text
Questions about fireplace Erato	Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol:

Comments

Position	Relevance	Comment Text
1	Good	I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions.
2	Bad	Thanks for the reply. Have you used any ekorecim?
3	Bad	I have no idea what that is. :shock:
4	Bad	Buhahaha :D These jokes become, when you write in the phone. I meant the ECO MODE : hug: by the way your fireplace when it reaches a certain temp stops work?
5	Good	The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes.

Question Subject		Question Text
Questions about fireplace Erato		Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol:
Comments		
Position	**Relevance**	**Comment Text**
1	Good	I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions.
2	Bad	Thanks for the reply. Have you used any ekorecim?
3	Bad	I have no idea what that is. :shock:
4	Bad	Buhahaha :D These jokes become, when you write in the phone. I meant the ECO MODE : hug: by the way your fireplace when it reaches a certain temp stops work?
5	Good	The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes.

15

How often to lit the fireplace? 🔍

1. Find similar question

Question Subject	Question Text
Questions about fireplace Erato	Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol:

Comments

Position	Relevance	Comment Text
1	Good	I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions.
5	Good	The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes.

2. Show useful responses

Content

What is CQA

Task Formulation

Method

Data

Experiments

Conclusion

CQA - Task Formulation

Given a question and comments from a forum thread,

CQA - Task Formulation

Given a question and comments from a forum thread,

order the comments with respect to their relevance with the question.

How often to lit the fireplace? 🔍

1. Find similar question

Question Subject		Question Text
Questions about fireplace Erato		Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol:
Comments		
Position	**Relevance**	**Comment Text**
1	Good	I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions.
5	Good	The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes.

2. Show useful responses

Why CQA

This year, the team of FMI of Sofia University participated in the SemEval competition - task for CQA in English.

Our system achieved the best results on the main subtask.

(Data in English from Qatar Living - a large forum in Qatar)

Goals of the Current Research

1. Existing pipeline for CQA for English - performs very well

2. Data from the largest forum in Bulgaria - BG Mamma

3. No research for CQA in Bulgarian so far

Goals of the Current Research

1. Existing pipeline for CQA for English - performs very well

2. Data from the largest forum in Bulgaria - BG Mamma

3. No research for CQA in Bulgarian so far

Would applying machine translation work well for CQA for Bulgarian?

Content

What is CQA
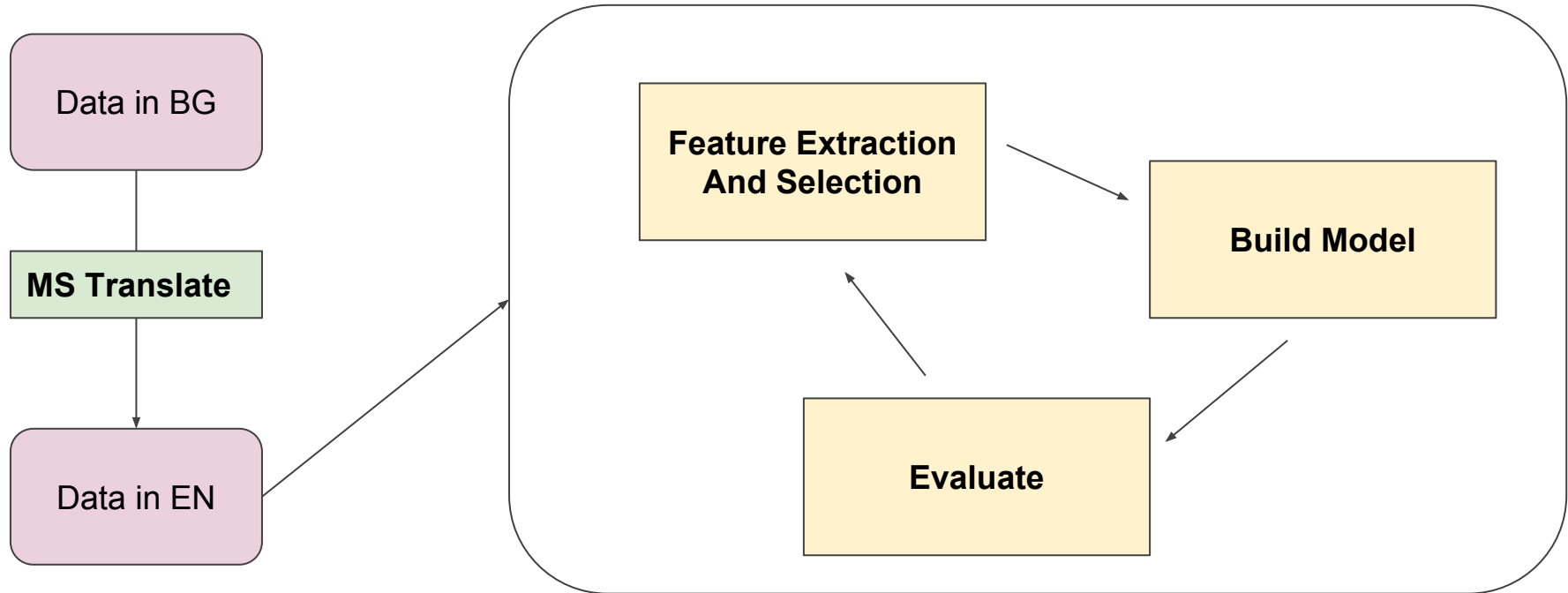
Task Formulation

Method

Data

Experiments

Conclusion

How it works

How do we know it works? (Evaluation)

Mean Average Precision (MAP)

How many good answers are put on the top positions?

The Good comments should be ordered 'over' the Bad comments.

$$MAP@5 = \sum_{q=1}^{Q} AveP(q)/Q$$

$$AveP@5 = \sum_{k=1}^{5} P(k)/5$$

Classifier

Classification instances: features extracted for question-answer pair.

SVM classifier with RBF kernel.

Use **probability for Good class** as ranking score.

Baseline

Use **1/comment position** as ranking score

MAP for baseline is ~70% on our test set

Goal: Improve this baseline with significant amount

Features

Metadata

Lexical

Semantic

Metadata Features

- Whether the comment is written by the author of the question
- Comment rank in the thread
- Ratio of the comment length to the question length (number of tokens)
- Number and order of comments from the same user in a particular thread
- Presence and the number of URLs in the question and in the comment

Lexical Features

- Question marks and question words in question and comment
- Smileys, "thank you", numbers
- Persons, Locations, Organizations, Addresses
- Level of readability and complexity of the text (Aluisio et al., 2010)
- POS count in question and comment
- Average number of words per sentence, type-to-token ratios
- Number of misspelled words and offensive words from a dictionary

Semantic Features

- **Topic Modeling** is used for training of 100 topics from questions and comments from the Qatar Living training data
- **Word Embeddings** trained with Word2Vec(Mikolov et al., 2013) on the Qatar Living forum data
- **Cosine distances** between the text of the question and of the comment (distance between the sum of the embeddings of all words in the question and in the comment text)

Content

What is CQA

Task Formulation

Method

Data

Experiments

Conclusion

Data

Data from BG-Mamma

Topics containing "въпрос"

First 5 answers from the thread (answers with position from 1 to 5)

Annotated the answers as "Good" or "Bad" in respect to the question.

Additional training set from Qatar Living forum (from SemEval 2016).

Data Stats

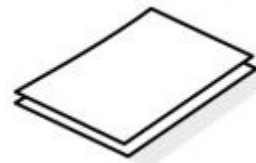|	Question-Answer pairs	Questions	Good Comments	Bad Comments
Training set:	250	50	84	166
Test set:	150	30	49	101
QL train set (SemEval Train 1) - first 5 comments	7055	1411	3021	4034

How do we overcome the shortage of training data?

Domain Adaptation (Daume III, 2007)

Domain Adaptation

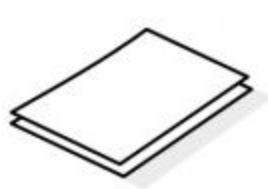Small annotated training set from one domain we need to classify

"target domain"

And large annotated set from another domain

"source domain"

Domain Adaptation

Combine the data so that we can use a larger annotated corpus for training.

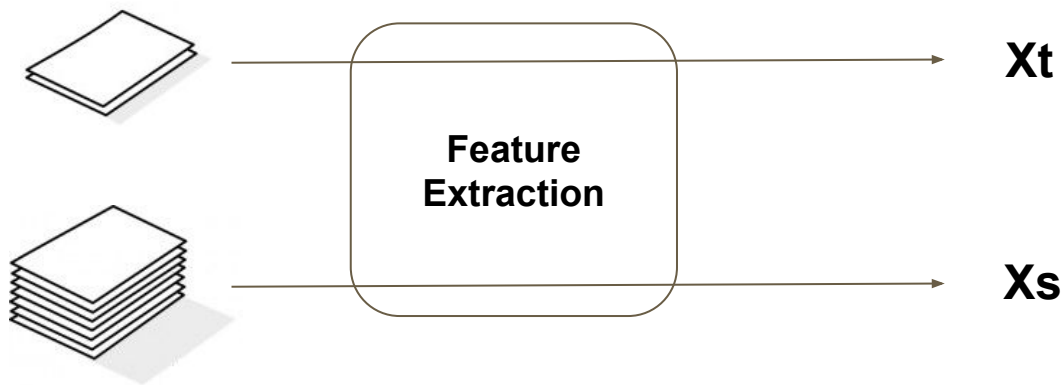Target Domain + Source Domain = Combined Training Set

Domain Adaptation

Extract features from source and target data

Feature Extraction

Xt

Xs

** **Xt** and **Xs** - feature vectors for target and source domain (the same features)*

Domain Adaptation

Extend feature space

Training Data:

- From Target Domain: **X = {Xt, 0, Xt}**
- From Source Domain: **X = {Xs, Xs, 0}**

Test Data:

- Only from Target Domain: **X = {Xt, 0, Xt}**

Domain Adaptation

Would that work?

Domain Adaptation

Would that work?

Yes, according to research. And to our experiments.

Content

What is CQA

Task Formulation

Method

Data
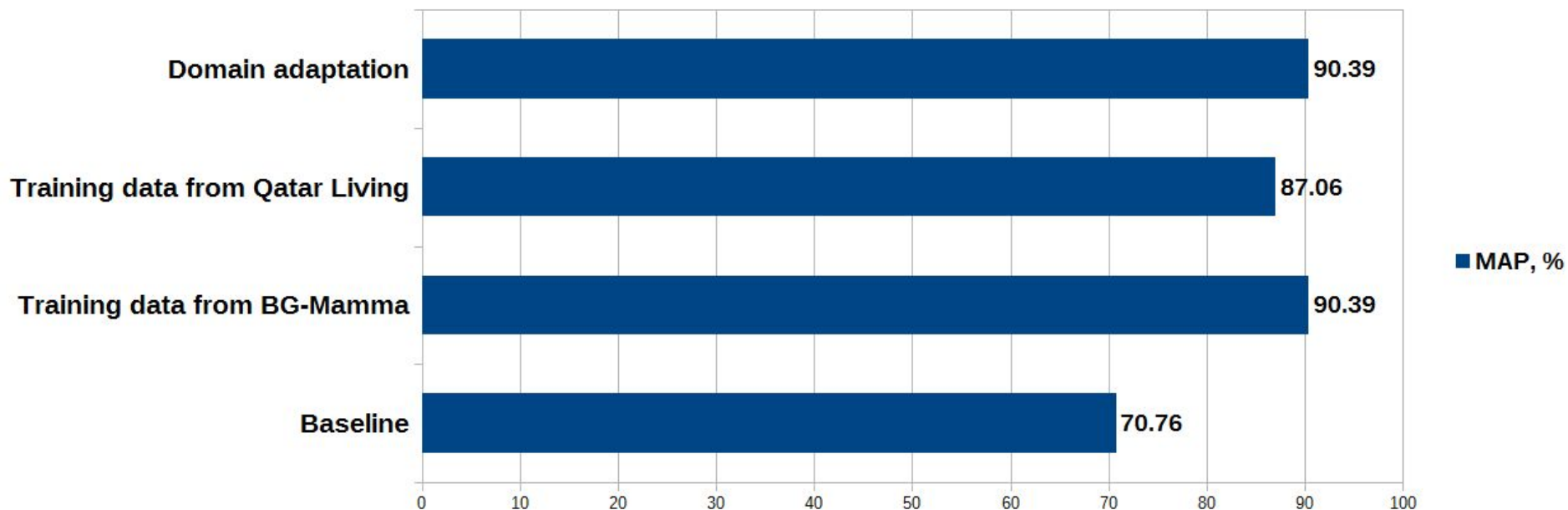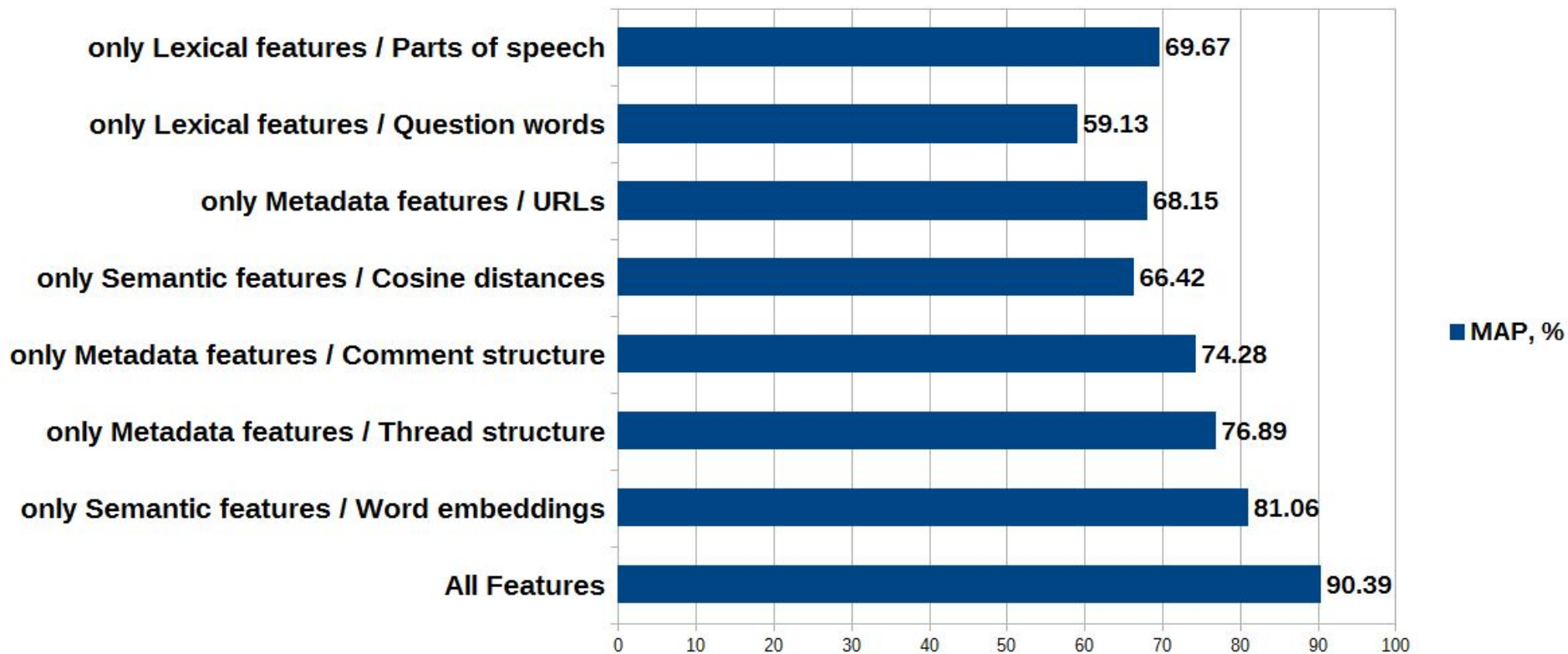
Experiments

Conclusion

Experiments - Different training sets on all features

	MAP, %
Domain adaptation	90.39
Training data from Qatar Living	87.06
Training data from BG-Mamma	90.39
Baseline	70.76

Experiments - only specific feature groups

Feature group	MAP, %
only Lexical features / Parts of speech	69.67
only Lexical features / Question words	59.13
only Metadata features / URLs	68.15
only Semantic features / Cosine distances	66.42
only Metadata features / Comment structure	74.28
only Metadata features / Thread structure	76.89
only Semantic features / Word embeddings	81.06
All Features	90.39

Trained on the data from BG-Mamma

Experiments - all features without some feat. groups

Feature group	MAP, %
All - Lexical features / Parts of speech	90.94
All - Lexical features / Question words	90.39
All - Metadata features / URLs	90.39
All - Semantic features / Cosine distances	88.72
All - Metadata features / Comment structure	90.11
All - Metadata features / Thread structure	85.83
All - Semantic features / Word embeddings	86.22
All Features	90.39
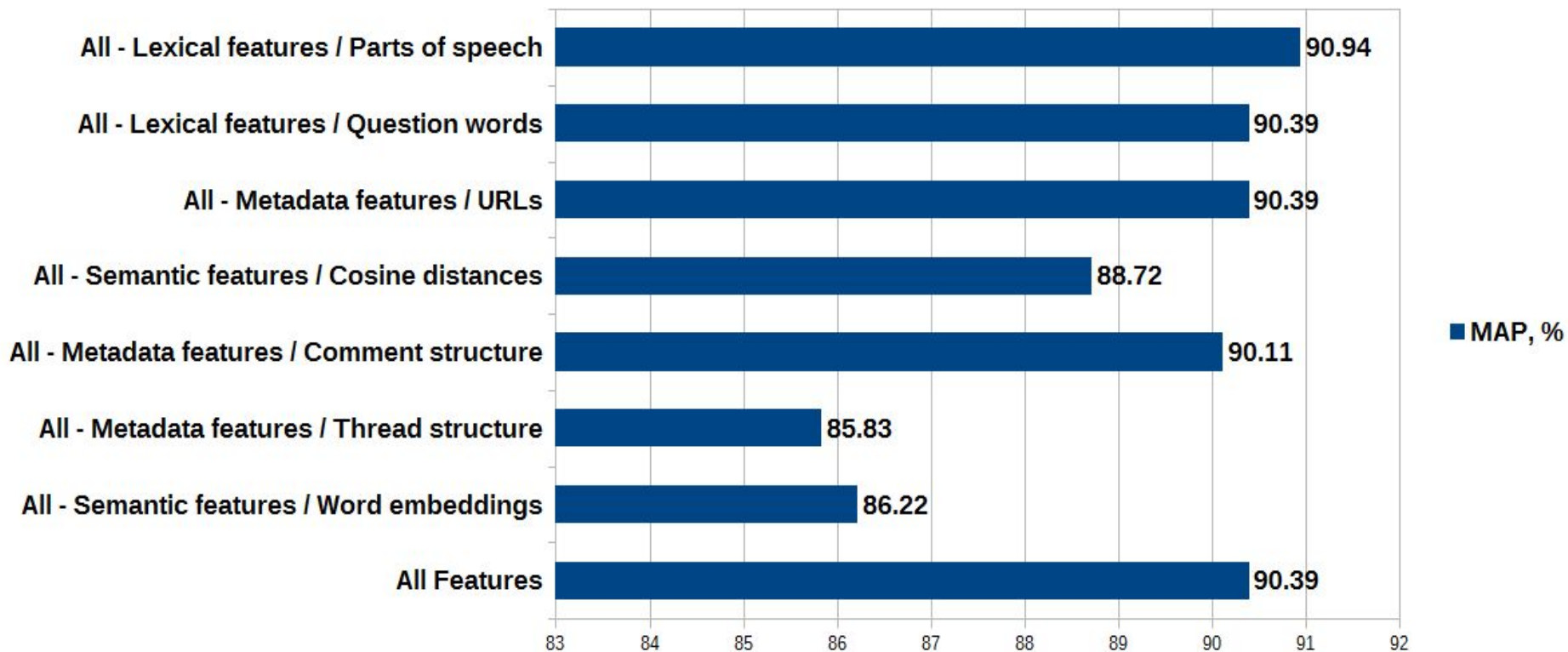
Trained on the data from BG-Mamma

Experiments - Results

Works **best** trained on data from **BG-Mamma** and **domain adaptation**

Best features:

- Word embeddings
- Metadata about the thread structure

Least significant features:

- Question words
- URLs - presence and count
- Parts of speech

Content

What is CQA

Task Formulation

Method

Data

Experiments

Conclusion

Conclusion

We presented our research for CQA for Bulgarian

1. The method of using machine translation from Bulgarian to English and using existing pipeline for English works well
2. Results are comparable to the ones tested in English
3. Achieved significant improvement over the baseline

Conclusion

We presented our research for CQA for Bulgarian

1. The method of using machine translation from Bulgarian to English and using existing pipeline for English works well
2. Results are comparable to the ones tested in English
3. Achieved significant improvement over the baseline

This approach can be used for further research in the field

Future Work

Add more features: user features, PMI, credibility

Annotate more data for BG

Publish the corpus for Bulgarian

Integrate into real-life application?

Thank you!