University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab

take[lab];

Frequently Asked Questions Retrieval for Croatian Based on Semantic Textual Similarity

Mladen Karan, Lovro Žmak, Jan Šnajder

Introduction

take[lab];

- Frequently Asked Questions (FAQ) databases are a popular way of getting domain-specific expert answers to user queries.
- A FAQ database consists of many question - answer pairs (FAQ pairs). In larger databases it can be difficult to manually find a relevant FAQ pair.
- Automated retrieval is challenging
  - Short texts cause keyword matching to perform poorly
- The goal of this work is to build a FAQ retrieval system for Croatian

Outline

take[lab];

- Data set
- Retrieval model
- Features
- Results
- Conclusion

Data set

take[lab];

- From the web we obtained the FAQ of Vip - a Croatian mobile phone operator (1222 unique FAQ pairs)
- Ten annotators were asked to create 12 queries each
- The annotators were then asked to paraphrase the queries
  - Turn into a multi sentence query
  - Change the syntax
  - Substitute some words with synonyms
  - Turn into declarative sentence
  - Combination of the above

Data set

take[lab];

- For each set of paraphrased queries we retrieve potentially relevant documents using a pooling method (including keyword search, phrase search, tf-idf and language modeling)
- The annotators were asked to review the retrieved set, assigning a binary relevance score to each retrieved FAQ pair. To reduce bias the pairs are presented in random order
- FAQ pairs not retrieved by the pooling method were assumed to be irrelevant

Data set

take[lab];

- The annotated data set includes:
  - A list of queries
  - A list of relevant FAQ pairs for each query
  - Additional metadata (i.e. categories of FAQ questions and information about annotators)
- The data set is freely available for research purposes (`takelab.fer.hr/data/faqir`)
- We focus only on queries which have at least one answer (327 of them)

Retrieval model

take[lab];

- We frame the FAQ retrieval task as a supervised machine learning problem.
- A classifier(SVM) is trained on annotated data:
  - Input – a query and a FAQ pair
  - Output – a binary relevance decision and a confidence score
- The classifier decision itself is not directly used, rather, the results are ordered by classifier confidence
- A variety of semantic similarity metrics is are used as features

Features – ngram overlap

take[lab];

- The coverage of text $T1$ with words from $T2$:

$$no(T1, T2) = \frac{|T_1 \cap T_2|}{|T_1|}$$

- The ngram overlap feature is the harmonic mean of $no(T1, T2)$ and $no(T2, T1)$
- It is calculated on unigrams and bigrams between the user query and both the FAQ question and FAQ answer

Features – ngram overlap

take[lab];

- To account for varying importance of words they can be weighted using information content (ic)
- The weighted coverage of text $T1$ with words from $T2$:

$$wno(T_1, T_2) = \frac{\sum_{w \in T_1 \cap T_2} ic(w)}{\sum_{w' \in T_1} ic(w')}$$

- The weighted ngram overlap feature is the harmonic mean of $wno(T1, T2)$ and $wno(T2, T1)$

Features – tf-idf

take[lab];

- Cosine similarity between query and FAQ pair bag-of-words vectors
- The elements of the vectors are weighted using tf-idf
- The FAQ pair is considered a single document (no distinction between the question and answer parts)

Features – LSA

take[lab];

- LSA derived word vectors ([Karan et al., 2012]) from the HrWaC corpus ([Ljubešić & Erjavec, 2011])
- The vector of a text T is derived compositionally ([Mitchell & Lapata, 2008]):

$$v(T) = \sum_{w \in T} v(w)$$

- The similarity of texts is given by the cosine of their vectors
- Computed between the user query and both the FAQ question and FAQ answer
- Weighted variant:

$$v(T) = \sum_{w_i \in T} ic(w_i) v(w_i)$$

Features – ALO

take[lab];

- Aligned lemma overlap ([Šarić et al., 2012])
- Given texts $T1$ and $T2$ greedily align words:
  - Find the most similar (LSA similarity) pair of words and remove them from futher consideration
  - Repeat until all there are no more words to pair up
- Calculate similarity for each pair ($ssim$ = LSA similarity)

$$sim(w_1, w_2) = \max(ic(w_1), ic(w_2)) \times ssim(w_1, w_2)$$

- Calculate the overall similarity

$$alo(T_1, T_2) = \frac{\sum_{(w_1,w_2)\in P} sim(w_1, w_2)}{\max(length(T_1), length(T_2))}$$

Features – QC

take[lab];

- Question classification data set containing 1300 questions ([Lombarović et al., 2011])
- Question classes: *numeric, entity, human, description, location, abbreviation*
- Using document frequency the most frequent 300 words and 600 bigrams are selected as features
- SVM - 80% accuracy
- The classifier outputs for the user query and FAQ question are included as features

Features – QED

take[lab];

- Query expansion dictionary
- Motivated by brief analysis of system errors. Aims to:
  - Mitigate minor spelling variances
  - Make similarity of cross-POS or domain specific words explicit
  - Introduce rudimentary world knowledge useful for the domain
- The dictionary includes a list of rules in the form *word - expansionword1, expansionword2, ...*
- In total there are 53 entries in the dictionary

Features – QED

take[lab];

Query expansion examples

Query word	Expansion words
face	facebook
ograničiti (*to limit*)	ograničenje (*limit*)
cijena (*price*)	trošak (*cost*), koštati (*to cost*)
inozemstvo (*abroad*)	roaming (*roaming*)
ADSL	internet

Evaluation

take[lab];

- Classifier performance is evaluated using the F1 measure
- FAQ retrieval system performance is evaluated using standard IR metrics:
  - Mean Reciprocal Rank (MRR)
  - Mean Average Precision (MAP)
  - R Precision (RP)
- All metrics are calculated using a 5 - fold cross validation over the 327 available user queries.
- A baseline FAQ retrieval system is based on tf-idf

Evaluation

take[lab];

Features used in the models

Feature	RM1	RM2	RM3	RM4	RM5
NGO	+	+	+	+	+
ICNGO	+	+	+	+	+
TFIDF	–	+	+	+	+
LSA	–	–	+	+	+
ICLSA	–	–	+	+	+
ALO	–	–	+	+	+
QED	–	–	–	+	+
QC	–	–	–	–	+

Results

take[lab];

Classification results

Model	P	R	F1
RM1	14.1	68.5	23.1
RM2	25.8	75.1	37.8
RM3	24.4	75.4	36.3
RM4	**25.7**	**77.7**	**38.2**
RM5	25.3	76.8	37.2

Results

take[lab];

FAQ retrieval results

Model	MRR	MAP	RP
Baseline	0.341	21.77	15.28
RM1	0.326	20.21	17.6
RM2	0.423	28.78	24.37
RM3	0.432	29.09	24.90
RM4	**0.479**	**33.42**	**28.74**
RM5	0.475	32.37	27.30

Results

take[lab];

- Most frequent causes of error:
  - Lexical interference – a non relevant FAQ pair can still have high lexical overlap
  - Lexical gap – lack of lexical overlap
  - Semantic gap – reasoning and/or world knowledge are required
  - Word matching errors – informal spelling variations

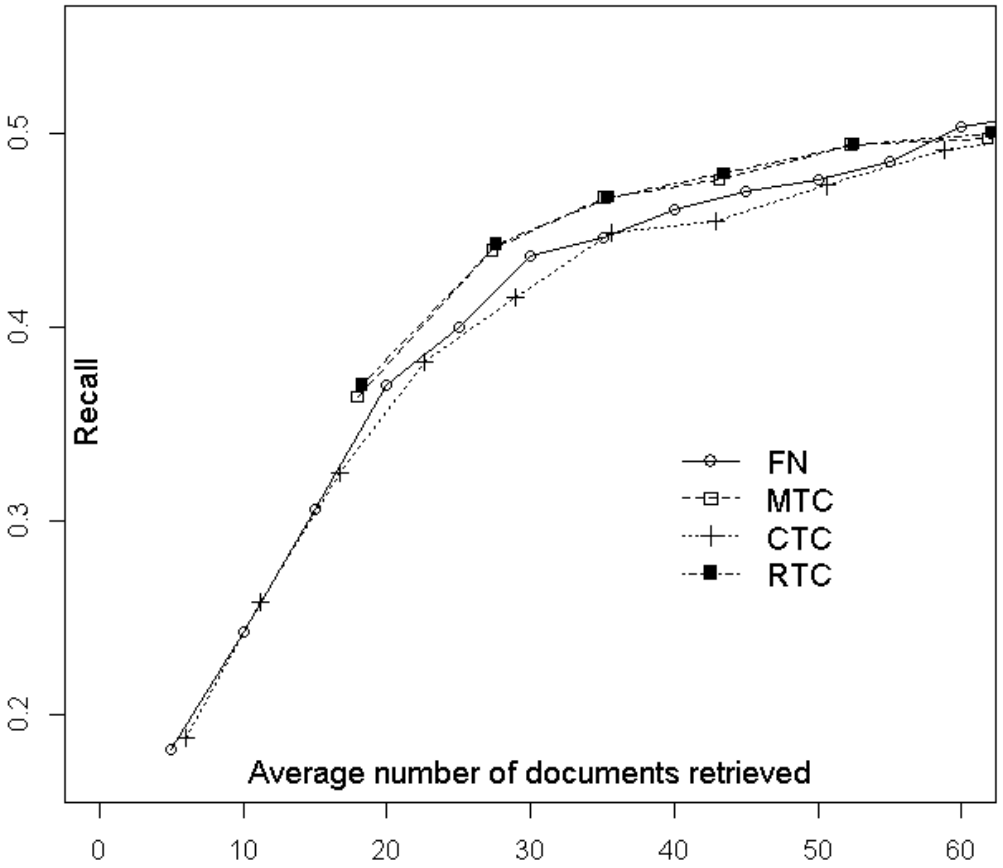
Results

take[lab];

- Presenting the entire ordered list puts an unnecessary burden on the user
- The list can be shortened using different cutoff criterions:
  - FN – first N
  - MTC – measure criterion
  - CTC – cumulative measure criterion
  - RTC – relative measure criterion
- A better criterion will yield higher recall with less retrieved documents

Results

take[lab];



Conclusion and future work

take[lab];

- A FAQ retrieval engine was built based on supervised machine learning using semantic similarity features
- Deceivingly high or low word overlap remains a problem, a possible solution is to use syntactic information
- The query expansion dictionary proved quite beneficial. The generation of expansion rules could be automated by analysing query logs collected over a longer time span ([Cui et al., 2002], [Kim & Seo, 2006])
- from a practical perspective, work on scaling up the system to large FAQ databases is required

References

Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2002).
Probabilistic query expansion using query logs.
In *Proceedings of the 11th international conference on World Wide Web* (pp. 325–332).: ACM.

Karan, M., Šnajder, J., & Dalbelo Bašić, B. (2012).
Distributional semantics approach to detecting synonyms in Croatian language.
In *Information Society 2012 - Eighth Language Technologies Conference* (pp. 111–116).

Kim, H. & Seo, J. (2006).
High-performance FAQ retrieval using an automatic clustering method of query logs.
Information processing & management, 42(3), 650–661.

Ljubešić, N. & Erjavec, T. (2011).
HrWaC and SlWaC: compiling web corpora for Croatian and Slovene.
In *Text, Speech and Dialogue* (pp. 395–402).: Springer.

Lombarović, T., Šnajder, J., & Bašić, B. D. (2011).
Question classification for a Croatian QA system.
In *Text, Speech and Dialogue* (pp. 403–410).: Springer.

Mitchell, J. & Lapata, M. (2008).
Vector-based models of semantic composition.
Proceedings of ACL-08: HLT, (pp. 236–244).

Šarić, F., Glavaš, G., Karan, M., Šnajder, J., & Bašić, B. D. (2012).
TakeLab: systems for measuring semantic text similarity.
In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 441–448).: Association for Computational Linguistics.

Thanks for your attention!

take[lab];

Text Analysis and Knowledge Engineering Lab
www.takelab.hr
info@takelab.hr, takelab@fer.hr