Towards Czech Answer Type Analysis

Daša Kušniráková and Marek Medveď

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xkusnir, xmedved1}@fi.muni.cz

Abstract. In this paper, we introduce two answer type detection systems for Czech language. Based on the input question, the goal of these tools is to recognise the question type and extract an appropriate answer type. Except for the same goal, these systems are completely different. The first one is a rule based system utilising Czech Wordnet for hypernym detection. The second one uses a machine learning approach in form of a neural network. We present architectures of these two systems and offer a detailed evaluation on more than 8,500 question-answer pairs using the SQAD v2.1 benchmark dataset.

Keywords: question answering; question classification; answer classification; Czech; Simple Question Answering Database; SQAD

1 Introduction

Open domain question answering (QA) systems have seen a great progress in recent years. Using neural networks models [1,2] and large datasets, e.g. SQuAD [3], the systems have become more and more usable.

The majority of QA tools consists of several modules that contribute to the final system performance. In this paper, we present answer type detection module that usually appears at the beginning of the processing pipeline mostly on the pre-processing level, whose main task is to determine the answer type according the input question. We introduce two implementations of such answer type detection tool. The first one is represented by a system based on rules enriched by a hypernymic dictionary, whereas the second one utilises a recurrent neural network model. Both systems will be tested inside the AQA system [4,5] pipeline and the answer type detection is expected to improve the decision process in the Answer extraction module of AQA (see Figure 1).

The following chapters provide a detailed specification of the rule based as well as the machine learning based system. In the last section, we offer a thorough evaluation of both systems, for which the benchmarking dataset SQAD v2.1 [6] database consisting of 8,566 questions-answer pairs has been used.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018, pp. 41–51, 2018.

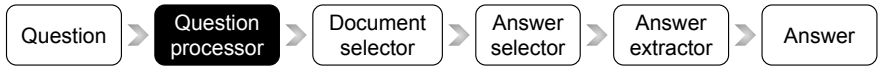

Fig. 1: AQA system visualisation

2 Question and Answer Type Detection

Same as for other classification issues, different approaches can be applied even when dealing with question and answer type detection. While some systems have been developed using rule based approach [7,8], machine learning based approach [9,10] has become more popular over the years. Because of the need to improve the overall performance of the AQA system, two systems have been developed, while each is based on a different approach. The rule based as well as the machine learning based system are described below.

2.1 Rule Based System

The rule based approach has been used from the early beginnings of dealing with QA type detection – e.g. a tool for question classification introduced in [11] capable of distinguishing between three classes. Even though the approach is on the decline these days due to more effective methods, rule based systems can still achieve satisfactory results. The systems introduced in [7,8] are able to classify required answer types with precision up to 83%.

The core of the developed rule based system introduced in this paper is formed by a set of hand-written rules approaching different features extracted from the question during the preprocessing phase. Such features include lexical features, POS tags, the *question keyword* and its hypernyms. The keyword is represented by the main (head) question meaning noun.

The keyword extraction algorithm is based on the following three rules:

- The question keyword candidate is the first noun after the relative pronoun *"který"* (which) or *"jaký"* (what), if such relative pronoun is present in the question and is not part of a relative sentence.
- Otherwise the first noun after the first verb in question is selected as the candidate for question keyword.
- The candidate becomes the final keyword unless it is one of words *"název"* (title), *"pojem"* (concept), *"termín"* (term), *"typ"* (type), *"část"* (part), or *"větev"* (branch). Otherwise the first following noun after the selected candidate is returned as the final keyword.

Keyword hypernyms are obtained by means of the Czech Wordnet API [12] in a two-step process. The Czech Wordnet is queried for the first time to find all

possible senses of the keyword extracted from the question and subsequently, the Wordnet API is queried again to create a list of hypernyms for three[1] most common word senses.

After all features have been obtained during the preprocessing phase, type detection rules are applied step by step to the question. If the rule's conditions are met by the question which is being classified, the appropriate labels representing question and answer types are returned. A schematic description of the QA type detection process is presented in Figure 2.

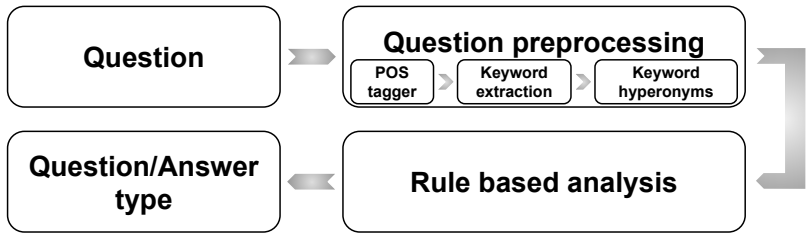

Fig. 2: Rule based question/answer type detection schema

The rules themselves are formed by any combination of the features recognized during the preprocessing phase. These include:

- keyword hypernym match:
  Example: `"<word>" in keyword.hypernym`
- important word recognition:
  Example: `"<word>" == words.lemma_at_index(0)`
  -> the first word in the sentence is the specified word
- question structure match:
  Example: `"k2" in words.tag_at_index(1)`
  -> the second word in the sentence is an adjective[2]

All pieces of information gained during the whole process of QA type detection (including the preprocessing as well as rule application phase) for a particular question can be seen in Figure 3.

[1] The number has been determined by testing of the overall performance of the system. Creating a list of hypernyms for both lower and higher number of word senses affects the performance in a negative way as the list becomes either too narrow or too broad, respectively.

[2] see [13] for more information about the POS tagset.

question: 'Jak se jmenovala první manželka Miloše Formana?'
(What was the name of the first wife of Miloš Forman?)
keyword: 'manželka' (wife)
hypernyms: ['manželka', 'jednotlivec', 'osoba', 'bytost', 'organismus']
(wife, individual, person, being, organism)
rule: `(PERSON; PERSON) -> "osoba" in keyword.hypernym`

Fig. 3: A question/answer type rule example: *if* `"osoba"` *(person) is one of the question's keyword hypernyms, then the question type is* `PERSON` *and the answer type is also* `PERSON`.

2.2 Machine Learning Based System

In comparison to the rule based approach, machine learning makes the process of question analysis and classification more automatic. Apart from that, these systems are able to achieve results comparable or even outperforming with other approaches. In the systems introduced in [9] or [10], the accuracy reaching for fine-grained classes around 90%, for coarse-grained classes even up to 95%.

In addition to the rule based system described in the previous section, a system for question and answer type detection based on machine learning – a Long Short-Term Memory (LSTM) network has been developed, too. Recurrent neural network has been chosen due to its ability to handle sequential input data of any length, while the LSTM unit is capable of dealing with the exploding and vanishing gradient problem.

Our proposed LSTM model consists of four layers. It contains two stacked LSTM layers with a dropout layer applied in between, and a linear layer. A visual representation of the model's architecture is shown in Figure 4. The two-layered LSTM architecture has outperformed both single and a three stacked LSTM layers by more than 80% and 5% in experimental attempts, respectively. The model has been trained with the usage of cross-entropy as loss function and with 40 epochs, batch size of 64, dropout rate of 0.5 and learning rate of 0.001 as its hyperparameters.
The process of question and answer classification is performed in the following steps:

- The input question is split into individual words, which are subsequently converted into dense, 100-dimensional vectors. The vectors are obtained from pre-trained Fasttext word embeddings trained on Czech corpora of more than 10 milliard words.
- Words in the form of 100D vectors represent the input to the LSTM model. They are processed one-by-one by LSTM layers. For each sequence, only the most recent timestep (affected by the previous ones) of the LSTM network is passed to the Linear layer.
- The Linear layer transforms the LSTM output to a vector of scores for each question and answer type combination, which is created by the cartesian

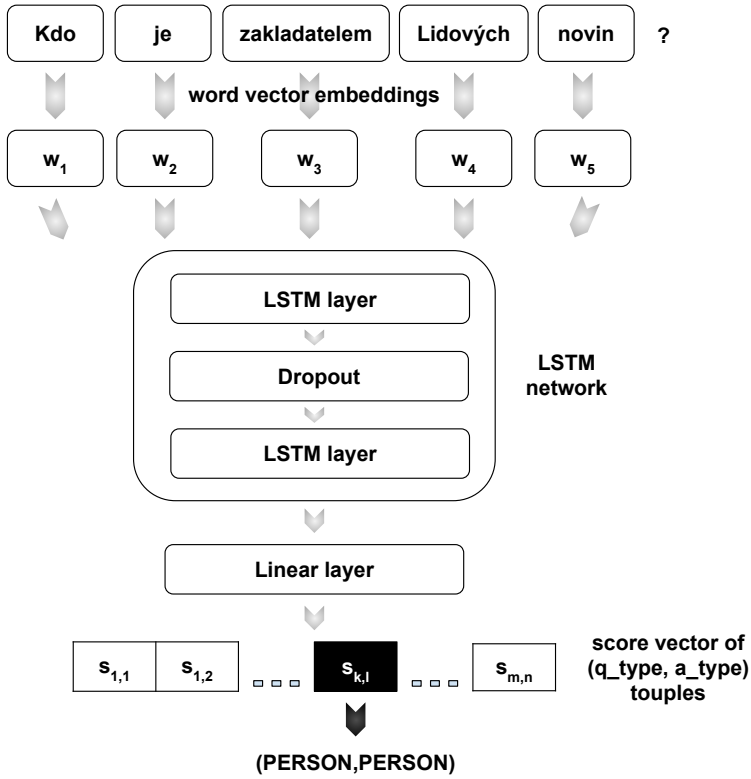

Fig. 4: LSTM network visualisation

product of all question and answer types. Considering there are possible 10 question and 10 answer classes, such vector created in the Linear layer is then 100-dimensional.

– The searched question and answer type is then determined from the position of the maximal score found in the vector returned from the Linear layer. The higher is the score, with the higher certainty is the particular combination of classes predicted by the model.

3 SQAD Database

The Czech Simple Question Answering Database, or SQAD [14,15], is a QA benchmarking dataset resource consisting of manually processed and manually annotated question-answer pairs. SQAD, originally created from Czech Wikipedia articles, now represents a consistent and representative data source for any model training and tool evaluation needs.

The SQAD v2.1 database currently contains 8,566 question-answer pairs which are related to the content of 3,149 Czech Wikipedia articles. The SQAD database is organised in structured records (one QA pair corresponds to one record) consisting of 6 items:

- the *question*,
- the *correct answer* (as can be extracted from the document),
- *answer selection* – the context of the correct answer, one or two sentences,
- the full *article text*
- the *source URL* in Wikipedia
- *question-answer metadata* containing *types of the question and the correct answer*.

All texts have been manually corrected and enriched by base word forms (lemma) and Part-Of-Speech (POS) annotation (DEsamb [16,13]).

The dataset contains annotation with classification of each record into categories for the question type and the actual correct answer type. The sets of possible types [15] took inspiration from the large benchmark dataset for English, the Stanford Question Answering Dataset [3].

The distribution of question classes over answer classes is displayed in Table 1.

Table 1: SQAD v2.1 distribution matrix of question and answer types

Q type /A type	PER.	DENOT.	ENT.	OTHER	ORG.	D./T.	LOC.	NUM.	ABB.	Y/N
PERSON	1,016	0	2	0	3	0	2	0	0	0
ENTITY	20	101	1,031	378	204	1	7	1	2	0
ADJ_P.	7	0	8	216	0	0	0	2	0	0
D./T.	0	0	1	2	0	1,844	0	4	0	0
LOC.	1	0	14	5	3	0	1,501	0	0	0
NUM.	1	0	0	2	0	0	0	910	0	0
ABB.	0	1	0	0	0	0	0	0	80	0
CLAUSE	1	0	27	205	6	0	1	1	0	0
VERB_P.	2	0	1	0	0	0	0	0	0	937
OTHER	2	0	1	11	0	0	0	0	0	1

4 Evaluation

This section offers an evaluation of the question-answer types detection on SQAD v2.1 database for both rule based and machine learing based systems. For correct evaluation, the database has been divided into parts. For the rule based system, the dataset has been split into training and testing set, for the LSTM network into training, evaluation and testing set. All parts are properly

balanced to maintain each question/answer type present in each division. The exact numbers of records in training, evaluation and testing sets for each system are listed in Table 2.

Table 2: Number of records after dataset splitting

	training	evaluation	testing
Rule based system	4,279	-	4,287
LSTM network	7,011	735	820

The final evaluation of the rule based system as well as the LSTM network is present in Table 3 and Table 4, respectively. The evaluation is calculated by weighted average process that is more suitable for multiclass classification setting.

Table 3: QA types detection evaluation – rule-based system

	precision	recall	F1
question t.	88.77%	87.79%	88.28%
answer t.	85.05%	84.52%	84.78%
both types	82.43%	82.93%	82.68%

Table 4: QA types detection evaluation – LSTM network

	precision	recall	F1
question t.	91.59%	90.73%	91.16%
answer t.	89.76%	89.14%	89.45%
both types	86.15%	87.07%	86.61%

4.1 Rule Based System

The recall of both types detection is **82.93%**, while the combined precision is 82.43% with question type precision of 88.77% and the answer type precision of 85.05%. The question type detection achieves recall of 87.79% and F1 measure going up to 88.28%. A detailed confusion matrix of all the expected and predicted question types is presented in Figure 5. According to the results, it can be seen that ENTITY class is among the most complex ones as entities can be expressed in several ways. A detailed evaluation of the answer type detection is present in Figure 6, where the most confusing classes for the system are ENTITY, OTHER and PERSON. This may call for further specification of the members of the OTHER class.

Table 5: Question type confusion matrix – rule-based system

	expected									
predicted	AB	APHR	CL	D/T	ENT	LOC	NUM	OTH	PER	VPHR
ABBR.	37	1	1	0	19	3	1	0	0	0
ADJ_P.	1	52	4	0	49	6	6	0	4	0
CLAUSE	1	0	35	0	14	4	0	0	5	0
D/T	0	0	1	916	16	0	2	0	1	1
ENTITY	0	44	71	3	685	41	13	2	40	8
LOC.	0	6	1	0	22	695	3	0	3	1
NUM.	1	4	1	4	8	0	422	0	0	0
OTHER	0	1	3	2	25	7	7	5	3	6
PERSON	0	8	3	0	33	6	2	0	455	0
V_PHR.	0	0	0	0	0	0	0	0	0	454

Table 6: Answer type confusion matrix – rule-based system

	expected									
predicted	AB	D/T	ENT	LOC	NUM	ORG	OTH	PER	DEN	Y/N
ABBR.	37	0	9	3	1	1	9	2	0	0
D/T	0	915	7	0	2	1	8	1	2	1
ENTITY	0	2	405	32	14	19	191	40	10	5
LOC.	0	0	7	693	3	9	15	3	0	1
NUM.	1	3	3	0	423	0	9	0	1	0
ORG.	1	0	30	5	0	61	24	6	0	0
OTHER	2	2	46	16	14	10	138	19	3	7
PERSON	0	0	12	7	2	13	18	452	3	0
DENOT.	0	0	1	1	1	1	3	0	38	0
YES_NO	0	0	0	0	0	0	0	0	0	454

4.2 LSTM Network

In the machine learning based system, the recall of both types is **87.07%** and the combined precision is 86.15%. The answer type precision is going up to 89.76%, while the question type detection achieves high precision going up to 91.59%. In general, it can be stated that the LSTM outperforms the rule based system by 2.7-5 points in each score according to the results. A detailed evaluation of question type detection is provided by Table 7. The deviation is most apparent for OTHER class, whose results have been affected by misclassifying the only record of this class. Table 8 presents the answer type detection results, where the most remarkable deficiencies can be seen namely in ENTITY, OTHER, and PERSON classes.

The LSTM network outperforms the rule based system according to the most recent results presented above even though no changes in hyperparameters of

the LSTM network have not been properly tested yet. The introduced LSTM system represents our first prototype so the architecture of the network may change in the near future to even better serve the classification task.

Table 7: Question type confusion matrix – LSTM network

	expected									
predicted	AB	APHR	CL	D/T	ENT	LOC	NUM	OTH	PER	VPHR
ABBR.	7	0	0	0	2	0	0	0	0	0
ADJ_P.	0	12	0	0	9	2	0	0	1	0
CLAUSE	0	0	9	0	12	0	0	0	3	0
D/T	0	0	0	175	0	0	0	0	0	0
ENTITY	0	5	6	0	129	3	1	0	9	1
LOC.	0	1	0	0	7	141	0	0	1	0
NUM.	1	1	0	1	0	0	87	0	0	0
OTHER	0	0	0	0	1	0	0	0	0	1
PERSON	0	0	0	0	6	0	0	0	95	0
V_PHR.	0	0	0	0	1	0	0	1	0	89

Table 8: Answer type confusion matrix – LSTM network

	expected									
predicted	AB	D/T	ENT	LOC	NUM	ORG	OTH	PER	DEN	Y/N
ABBR.	7	0	0	0	0	1	1	0	0	0
D/T	0	175	0	0	0	0	0	0	0	0
ENTITY	0	1	72	4	0	2	13	3	1	0
LOC.	0	0	2	140	0	2	2	1	0	0
NUM.	1	1	0	0	87	0	1	0	0	0
ORG.	0	0	1	0	0	12	1	3	0	0
OTHER	0	3	20	2	1	0	44	7	2	2
PERSON	0	0	3	0	0	2	2	96	0	0
DENOT.	0	0	2	0	0	0	1	0	9	0
YES_NO	0	0	0	0	0	0	1	0	0	89

5 Conclusion and Future Work

In this paper, we have introduced two different tools for question and expected answer type detection used in the Question processor and Answer extraction

modules of the question answering system AQA – a rule based system and a Long Short-Term Memory (LSTM) network.

The detection of the rule based system is based on a set of hand-written rules which determine QA types according to lexical, syntactic and semantic features obtained by the question processing. The module was trained on a balanced half of the SQAD questions and evaluated with the testing set of comparatively the same size. The resulting precision was 88.77% for question and 85.05% for answer types with the respective recall of 87.79% and 84.52%. The combined overall F1 measure was 82.68%.

The LSTM network is machine learning based and utilises a recurrent neural network model using Fasttext word embedding vectors. The model has been trained on 50% of the SQAD questions while next 10% have been used for model evaluation during the training phase and 40% for testing. The combined recall of both types is 87.07% with the question and answer type precision going up to 91.59% and 89.76% respectively. The results show that the LSTM system outperforms the rule based model by 2.7-5 points in each score.

The introduced question and answer type detection tools have been developed in order to improve the performance of the question answering system AQA. Because of the fact machine learning based systems have better presumptions for the future, it is planed to continue in the development of the LSTM model, which includes experimenting with its architecture and setting of hyperparameters.

Acknowledgements. This work has been partly supported by the Czech Science Foundation under the project GA18-23891S.

References

1. Wang, W., Yan, M., Wu, C.: Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Volume 1. (2018) 1705–1714
2. Hu, M., Peng, Y., Huang, Z., Yang, N., Zhou, M., et al.: Read + Verify: Machine reading comprehension with unanswerable questions. arXiv preprint arXiv:1808.05759 (2018)
3. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Association for Computational Linguistics (2016) 2383–2392
4. Medveď, M., Horák, A.: AQA: Automatic Question Answering System for Czech. In Sojka, P., et al., eds.: Text, Speech, and Dialogue 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings, Switzerland, Springer International Publishing (2016) 270–278
5. Medveď, M., Horák, A.: Sentence and Word Embedding Employed in Open Question-Answering. In: Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018), Setúbal, Portugal, SCITEPRESS - Science and Technology Publications (2018) 486–492

6. Šulganová, T., Medveď, M., Horák, A.: Enlargement of the Czech Question-Answering Dataset to SQAD v2.0. In Aleš Horák, Pavel Rychlý, A.R., ed.: Proceedings of the Eleventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, Brno, Tribun EU (2017) 79–84
7. Przybyła, P.: Question Analysis for Polish Question Answering. In: 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop, Association for Computational Linguistics (2013) 96–102
8. Laokulrat, N.: A survey on question classification techniques for question answering. KMITL Information Technology Journal **2**(1) (2013)
9. Silva, J., Coheur, L., Mendes, A.C., Wichert, A.: From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review **35**(2) (2011) 137–154
10. Krishnan, V., Das, S., Chakrabarti, S.: Enhanced Answer Type Inference from Questions Using Sequential Models. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05, Association for Computational Linguistics (2005) 315–322
11. Kwok, C., Etzioni, O., Weld, D.S.: Scaling question answering to the web. arXiv preprint arXiv:1808.05759 **19**(3) (2001) 2383–2392
12. Rambousek, A., Pala, K., Tukačová, S.: Overview and Future of Czech Wordnet. In McCrae, J.P., Bond, F., Buitelaar, P., Cimiano, P., 4, T.D., Gracia, J., Kernerman, I., Ponsoda, E.M., Ordan, N., Piasecki, M., eds.: LDK Workshops: OntoLex, TIAD and Challenges for Wordnets, Galway, Ireland, CEUR-WS.org (2017) 146–151
13. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. (2009) 13–16
14. Horák, A., Medveď, M.: SQAD: Simple Question Answering Database. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, Brno, Tribun EU (2014) 121–128
15. Šulganová, T., Medveď, M., Horák, A.: Enlargement of the Czech Question-Answering Dataset to SQAD v2.0. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017. (2017) 79–84
16. Šmerk, Pavel: K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech). PhD thesis, Faculty of Informatics, Masaryk University (2010)