Samsung Research Poland (SRPOL) at SemEval-2022 Task 9: Hybrid Question Answering Using Semantic Roles

Tomasz Dryjański[1], **Monika Zaleska**[1], **Bartłomiej Kuźma**[1], **Artur Błażejewski**[1], **Zuzanna Bordzicka**[1], **Klaudia Firląg**[1], **Christian Goltz**[1], **Maciej Grabowski**[1], **Jakub Jończyk**[1], **Grzegorz Kłosiński**[1], **Natalia Paszkiewicz**[1], **Bartłomiej Paziewski**[1], **Jarosław Piersa**[1], **Paweł Bujnowski**[1], **Piotr Andruszkiewicz**[1,2]

{t.dryjanski; m.zaleska; b.kuzma; a.blazejewsk; z.bordzicka; k.firlag; c.goltz; m.grabowski2; j.jonczyk; g.klosinski; n.paszkiewic; b.paziewski; j.piersa; p.bujnowski; p.andruszki2}@samsung.com

[1] Samsung Research Poland, Warsaw

[2] Warsaw University of Technology, Warsaw

Abstract

In this work we present an overview of our winning system for the R2VQ — Competence-based Multimodal Question Answering task, with the final exact match score of 92.53%. The task is structured as question-answer pairs, querying how well a system is capable of competence-based comprehension of recipes. We propose a hybrid of a rule-based system, Question Answering Transformer, and a neural classifier for N/A answers recognition. The rule-based system focuses on intent identification, data extraction and response generation.

1 Introduction

The goal of the task[1] was to develop a system applying existing knowledge to new situations, demonstrating a kind of understanding of a real-world domain. The competition presents a QA[2] challenge requiring linguistic and cognitive competencies that humans have while speaking and reasoning (Tu et al., 2022).

The task dataset contains questions belonging to "question families" based on CLEVR (Johnson et al., 2016), reflecting specific reasoning competences. These families were explicitly marked as 19 categories, the last one having no answer (N/A), but direct reference to these categories was prohibited by the task requirements.

The cooking recipes included in the dataset were provided with exceptionally extensive annotations containing semantic information. The authors applied CRL and span-based SRL using VerbAtlas (Di Fabio et al., 2019) for the reference inventory of frames and semantic roles. Subsequently, human annotators were asked to validate and correct frames and argument labels.

The dataset was split into training (26,526), validation (3,829) and test set (3,442 questions). At the competition evaluation stage, the answers to the latter were not revealed, but the annotations were retained.

Our source code is available on GitHub[3].

2 Related Work

In the recent years, deep learning systems trained on large datasets began to outperform humans and other algorithms in the whole QA discipline. Challenges presented in works such as SQuAD (Rajpurkar et al., 2018), MS MARCO (Nguyen et al., 2016), CoQA (Reddy et al., 2019), multilingual MLQA (Lewis et al., 2020) and others popularized various machine learning models for extractive QA. Meanwhile, visual and multimodal QA contests started to appear, e.g. VQA (Antol et al., 2015) or Audio-Visual Scene-Aware Dialog (Alamri et al., 2019). They require understanding of images, natural language and their mutual relations to produce answers. One should not overlook QA systems applying SRL annotations used in advanced answer and question generation, such as Fitzgerald et al. (2018).

QA systems were proposed for open domains as well as specific ones, including medicine, education, tourism, weather forecasting, etc. One of the most popular yet challenging topics for QA is cooking. The system in Khilji et al. (2021) required preparing a cooking-related ontology, categorizing questions and extracting potential answers.

[1] https://competitions.codalab.org/competitions/34056 (access Apr 28th, 2022).

[2] Abbreviations used in the text: QA: Question Answering; SRL: Semantic Role Labeling (or Labels); CRL: Cooking Role Labeling; EM: Exact Match; RC: Reading Comprehension; DNN: Deep Neural Networks.

[3] https://github.com/samsungnlp/semeval2022-task9

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1263 - 1273

Haussmann et al. (2019) focused on a knowledge graph used to answer a range of questions related to healthy diet.

Furthermore, food recognition could be perceived as a part or a preliminary step for cooking QA. Mohanty et al. (2021) and Akhi et al. (2018) seek for deep learning classifiers to properly identify food from real images.

3 System Overview

3.1 Questions Categorization

Because using original question categories was not allowed, we started with building a categorization solution. We based it on syntactic and lexical structure of the questions and used regular expressions as a way of distinguishing them; details can be found in Appendix C. Subsequently, to discover relationships between resulting question groups, as well as within them, we took SRLs and CRLs into account. They allowed us to determine a word or a phrase that should be included in the answer to a given question. Finally, we distinguished 17 question categories. To match the answers more effectively, some of them were later divided into subcategories:

1. COUNTING TIMES — counting how many times a given TOOL or HABITAT is used.
2. COUNTING USES — counting how many TOOLS or HABITATS are used.
3. COUNTING ACTIONS — counting how many actions it takes to do something.
4. ELLIPSIS — searching for direct object(s) which has undergone a certain process.
5. LOCATION (CRL) — searching for the place to which something is being transferred or in which it is located (a CRL is returned).
6. LOCATION (SRL) — similar to the above, but an SRL is returned.
7. METHOD — searching for a way of performing an action, with four subcategories according to which a CRL or an SRL is returned as an answer:
   - Question about a TOOL,
   - Question about an INSTRUMENT — objects or forces (such as heat, cold) that come in contact with an object and cause a change in it,
   - Question about an ATTRIBUTE — a property that a direct or indirect object possesses,
   - Question about a GOAL — the point to which something (e.g. temperature/heat/flame, consistency, thickness) needs to be brought.
8. LIFESPAN (HOW) — searching for a result of a process; a related action and its objects are returned as the answer.
9. LIFESPAN (WHAT) — similar to the above, but only related objects are inserted into the answer (without the action).
10. EVENT ORDERING — checking which action should be performed first.
11. RESULT — searching for expressions determining to what point a condition has changed.
12. TIME — searching for a specific expression relating to time.
13. EXTENT — searching for expressions specifying the range or degree of change.
14. PURPOSE — searching for expressions describing why an action needs to be performed.
15. CO-PATIENT — searching for indirect objects that undergo a process, are affected in a certain way, are situated in a particular location or are transferred to a different location.
16. SOURCE — searching for a starting point of a motion.
17. LOCATION CHANGE — searching for previous location of an object.

3.2 Approach Based on Semantic Roles

The system uses the following three-step path to find the answer: intent identification, data extraction and response generation.

Having the intent predicted, the question is dispatched to one of the per-category handlers. We designed the system to use a separate answerer for each category. LOCATION, RESULT, TIME, EXTENT, PURPOSE, CO-PATIENT and SOURCE share the same code after some parametrization, see Table 4. For the remaining categories (COUNTING, ELLIPSIS, LOCATION CHANGE, EVENT ORDERING, METHOD and both LIFESPANS) we use separate sub-engines, as we need to perform diverse tasks.

The implementations (except for METHOD) are pretty straightforward and obey the general rule:

- identify a reference verb and / or object in the question,
- search for a relevant sentence using the same verb / object in the given role (category-dependent) and extract relevant informa-

tion from the sentence using semantic roles (category-dependent again),
- if necessary, rephrase the information to form the answer.

Since METHOD contains four original categories (2, 6, 10 and 14) and direct use of the category ID was prohibited, the sub-engine for METHOD runs the above steps multiple times for: ATTRIBUTE, INSTRUMENT, GOAL and TOOL, returning the first found answer. The exact order of the labels was found empirically, by minimizing the number of category mismatches on the validation set.

In following sections we discuss the details.

Intent Identification

In almost every category the key to answering a question is identifying the verb and the object associated to it (jointly referred to as *intent*) and then finding the answer in the annotation.

First, a question classifier (see Section 3.1) is used to assign the question to the relevant category. Then, the analysis of the recipe is performed in an iterative way. We start with a small chunk (sentence or paragraph) to prevent mismatches resulting from looking too broadly. Verbs from the analyzed part are collected using either SRL (more specifically, tokens labeled as B-V), or a CRL and SRL combination, namely finding B-EVENT (CRL) with corresponding SRL (I-V or D-V). A detailed description of the annotation system is presented in Tu et al. (2022).

The next step of the intent identification requires iteration over the collected verbs to find a related object for each of them. The objects may be annotated in numerous ways:

- using SRL (e.g., PATIENT, THEME)
- using CRL (e.g., TOOL, HABITAT, EXPLICIT-INGREDIENT)
- using HIDDEN ROLES (e.g. DROP, HABITAT)

When both the verb and the associated object occur in the question, the system is ready to utilize this information to search for the answer in the recipe. More details of our algorithm can be found in Appendix C.

Data Extraction

Depending on the identified intent, the answer may appear in the passage either explicitly or implicitly. In the first case, the data essential for generating the answer is a direct span from the recipe and the system only needs to find an appropriate SRL and return it as the answer. However, for question categories where the answer is not explicitly mentioned in the passage, the process of data extraction is far more complicated. It requires calculating the actions, tracking object position, or collecting parts of the answer using all the information available in the annotation part: SRL, CRL, HIDDEN ROLES and the relations between them.

Response Generation

Generation of the final response is category-specific. In some cases, the gold answer contains only words annotated as a specific SRL (e.g. LOCATION, TIME). In other categories, the gold response contains the verb and the object from the question. There are categories where the system has to count occurrences of the object and return the number, as well as ones where the phrase *by using* is required at the beginning. See Appendix C for details.

3.3 DNN-Based Systems

QA is a well-established NLP task, mainly thanks to the advancements in attention-based DNN models. Thanks to fine-tuning, pre-trained BERT (Devlin et al., 2018) and its successors may be employed for downstream tasks, such as RC.

To examine how successful RC models could be for the competition, we tested the following ones: BERT, RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2020) in the extractive setup, i.e. taking text spans as predicted answers. We fine-tuned them on the SQuAD dataset and then on the task recipes from the train set. We used large versions of the models, and trained them for 5 (BERT), 12 (RoBERTa) and 15 (ELECTRA) epochs. The other hyperparameters used were: batch size: 8, learning rate: 1.5e-5, and max token sequence length: 512. Notably, we did not use provided annotations so that the solution was based solely on raw recipe texts.

N/A Classifier

An important part of the task is the correct identification of N/A answers in the QA pairs. The dataset contains about 9% of such, spread across all categories quite evenly. In most cases, the rule-based system was enough to identify the missing response in a cooking recipe. For more problematic situations we reached the best classification results by fine-tuning the *bert-base-multilingual-uncased*

model (Wolf et al., 2020) taken from the PyTorch Hugging Face repository[4].

4 Results

Our end-to-end hybrid system reached EM scores between 80% and 100% per question category and the official result overall amounted to 92.53%. Details can be found in Table 1.

Our pipeline starts with the semantic-based system. If no answer is returned, the RC system is used if its confidence threshold exceeds 98%. This fallback mechanism produced any significant improvement only for the LOCATION (SRL) category. We additionally consider the N/A Classifier result: if it exceeds the 99% certainty threshold, the N/A answer is returned. This operation enhances the results for the EVENT ORDERING category. Adding the DNN systems in such a way leads to a 0.145 pp result increase.

In the post-evaluation phase we made further improvements, mainly in the rule-based system, and we ultimately reached the 92.969% EM result.

Human annotators reached notably low EM score of 52%. It is mainly due to the fact that the exact match metric leaves no room for human creativity. A manual review of the semantic validity of the responses gave us 73% alignment with the gold answers. This is discussed further in Section 4.2.

Manual analysis revealed that only some elements of the images associated with the recipes relate directly to recipe content. This was also mentioned by task organizers. Only 62 from 500 analyzed pictures were considered helpful in answering questions by our human evaluators. They also reported that they were often assigned to a different recipe step. For these reasons we further disregarded the images and focused only on the textual part of the data.

4.1 DNN-based Systems

Table 2 shows RC models comparison. As expected, the best Exact Match score was achieved by ELECTRA, which is currently the top-performing model on the SQuAD benchmark.

Table 3 presents the percentage of test set questions that could be answered by an oracle extractive answering system, i.e. where the answer either can be found as a span from the recipe text or it is N/A. Such examples cover 35% of the test set, meaning that this is the upper limit for any extractive QA system. The result achieved by ELECTRA (EM equal to 31%) is in line with this estimation. Another 34.6% EM could be achieved with an extractive QA system by using additional post-processing.

This leaves out 30.4% examples, mainly from categories COUNTING, LIFESPAN and EVENT ORDERING, that require non-trivial processing (e.g. rephrasing) and/or aggregation of information from various parts of the recipe.

Based on these results we claim that ELECTRA or other BERT-based systems can be considered applicable for this type of task, yet they should be able to generate answers beyond plain span extraction. It would require improvements, such as making use of a generative model, feeding the semantic annotations along with recipe texts, and perhaps adjusting the models to specific question categories.

The N/A Classifier worked better when fed with the full recipe passage (i.e. *Ingredients* and *Directions*; $F1 = 82.7\%$). If provided only with *Directions*, the result dropped to $F1 = 76.6\%$. It shows that the *Ingredients* information plays an important role in the solution.

4.2 Human benchmark

The aim of creating the human benchmark described in this subsection has no other purpose but to measure to what extent our results (as well as the gold answers) are close to human reasoning, i.e. to the answers provided by an actual person. We did so as we did not find any information on human performance in materials provided by the organizers.

We asked a group of six linguists to answer 2,000 questions selected randomly from the validation set. We maintained similar percentages of each question category for the sample to be representative. Before starting their task, the linguists had become familiar with the train dataset to grasp the main idea and the structure of questions and answers. Importantly, they did not have access to the annotation so that they based their answers solely on the recipe texts and related pictures. We decided to take this approach assuming that the semantic annotation of the recipes serves as a partial equivalent of the general knowledge that AI lacks.

As already mentioned, the manual review of the human answers revealed that 73% of them have the same meaning as the gold answers. Other re-

[4] https://huggingface.co/models (access Feb 20th, 2022).

Table 1: Exact Match percentage per category. For the training (**Train**), validation (**Val**) and test (**Test**) sets we present the results of our hybrid system. **Post-Eval** shows our final post-evaluation results if different than **Test**. **Size** is a percentage of the given question category in the whole validation set. **Human** results were calculated based on a sample from the validation set as described in Section 4.2. For **Electra** we took the full test set.

Category	Size	Train	Val	Test	Post-Eval	Human	Electra
COUNTING TIMES	2.3	80.6	95.3	88.5		41.9	9.0
COUNTING ACTIONS	6.2	89.7	88.4	87.8		52.7	8.9
COUNTING USES	5.4	98.1	97.5	98.4		77.1	10.2
ELLIPSIS	13.8	89.2	89.3	89.5		20.9	22.7
LOCATION (CRL)	9.4	98.4	97.5	98.4		51.0	47.2
LOCATION (SRL)	8.0	95.6	96.5	95.3		69.5	80.1
METHOD	13.4	86.4	87.9	87.0	88.0	37.1	23.9
LIFESPAN (HOW)	5.4	89.1	91.6	88.7		5.1	10.8
LIFESPAN (WHAT)	5.1	93.7	93.9	92.6		15.6	21.1
EVENT ORDERING	15.8	97.1	97.8	96.7	97.2	93.4	9.8
RESULT	2.5	95.9	97.9	96.5		96.2	83.5
TIME	3.0	87.8	94.2	90.3		74.7	73.8
EXTENT	0.3	100.0	100.0	88.9		0.0	88.9
PURPOSE	1.2	98.2	100.0	97.6		81.8	82.9
CO-PATIENT	0.6	88 .4	95.8	85.0		64.3	90.0
SOURCE	0.6	96.4	100.0	100.0		68.4	31.0
LOCATION CHANGE	7.2	93.9	97.0	91.5	93.9	40.8	40
Total		**92.7**	**93.9**	**92.5**	**93.0**	**52.0**	**31.0**

Table 2: Reading Comprehension models results for the test set (0 - 100 range). **EM** — Exact Match score.

Model	F1	EM
BERT	36.9	30.7
RoBERTa	37.7	30.7
ELECTRA	**38.5**	**31.0**

sponses are semantically close, yet not identical. However, they often differ lexically from gold answers, resulting in the low overall EM score.

It is particularly visible in the ELLIPSIS category:

> **Question:** What should be tossed?
> **Gold answer:** the rice mixture and yogurt mixture
> **Human answer:** yogurt, sour cream, mustard, sugar, salt, pepper and rice mixture

The linguists also failed to return the gold answer when the question itself was semantically ambiguous. It was mostly applicable to the METHOD category and to both LIFESPAN categories. In the former, it results from various possible ways of understanding the English word *how*:

> **Question:** How do you slice the tomatoes?
> **Gold answer:** by using a knife
> **Human answer:** slice the tomatoes thinly

The LIFESPAN questions require listing ingredients needed to obtain something. The discrepancies between the gold answer and the human answer often resulted from a different nouns ordering or using a synonym:

> **Question:** How did you get the hot chocolate?
> **Gold answer:** by mixing the hot water, milk and mixture in the mug
> **Human answer:** by mixing the mixture with hot water or milk in a mug

As linguists did not see the annotation, their proposals were often different from the gold answer in categories where it was taken from SRL or CRL, such as METHOD subcategory concerning tools or habitats, or the COUNTING categories. Moreover, e.g. in both LIFESPAN categories, gold answers either listed the ingredients explicitly or returned the DROP value (one of the HIDDEN ROLES), such as "mixture", "soup" or "dough". From the human point of view those two kinds of responses would be equally correct:

> **Question:** What's in the mixture?

Table 3: Extractive Answering usability on the task. **EA** — answers present in source texts as non-empty spans. **NA** — N/A answers in the test set. **AQ** — Answerable Questions (EA + NA); i.e. an oracle system result. **EM** — Exact Match ($\leq$ AQ) actually achieved by our ELECTRA system. All results are provided as percentage and based on the test set.

Category	EA	NA	AQ	EM
COUNTING TIMES	0	9	9	9
COUNTING ACTIONS	0	9	9	9
COUNTING USES	0	10	10	10
ELLIPSIS	12	10	22	21
LOCATION (CRL)	50	10	60	47
LOCATION (SRL)	75	9	84	81
METHOD	13	12	25	24
LIFESPAN (HOW)	0	11	11	11
LIFESPAN (WHAT)	16	11	27	21
EVENT ORDERING	0	10	10	10
RESULT	79	5	84	83
TIME	67	13	80	74
EXTENT	78	11	89	89
PURPOSE	78	10	88	83
CO-PATIENT	80	10	90	90
SOURCE	81	5	86	31
LOCATION CHANGE	48	14	62	40
Total	25	10	35	31

Gold answer: the egg and mixture

Human answer: the butter, sugar, tangerine zest, vanilla, baking powder, salt and egg

The linguists obtained the best results in the EVENT ORDERING, RESULT, TIME and PURPOSE categories. Apart from the last one, those are closed-form questions that leave little room for semantic ambiguity.

We treated human benchmark as an interesting experiment that confirmed two hypotheses we had. Firstly, the answers provided by our model are often semantically close to the gold answers, as stated above. The scoring criteria reject any answer that is not identical to the gold one, which leads to allegedly poor human performance and makes the answer post-processing a daunting but crucial step. Secondly, there are some patterns in the task data that are remote from human thinking. The result of the experiment did not affect the final score — it served solely for analytic purposes.

5 Conclusion and future work

Our main contribution is the hybrid system for the cooking-related QA. While we are satisfied with the result, the ~7% error rate still leaves some room for improvement.

The most challenging task for our system was the correct intent identification. This is visible in the fairly low results in the METHOD category. It may relate to four different intents, and we did not always distinguish them properly. Other problematic aspects were counting actions and objects and generating answers that contain all required items in the right order. These issues solely contribute to as much as 5.5 out of 7 pp constituting the whole the error rate.

The obvious question left unanswered is the possibility of SRL/CRL annotation automation, also for other competence domains. This is a missing component in a full end-to-end application of our solution.

References

Amatul Akhi, Farzana Akter, Tania Khatun, Mohammad Uddin, Mohammad, and Shorif Uddin. 2018. Recognition and Classification of Fast Food Images. *Global Journal of Computer Science and Technology*, 18.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: A novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong

Kong, China. Association for Computational Linguistics.

Nicholas Fitzgerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Steven Haussmann, Oshani Wasana Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. 2019. Foodkg: A semantics-driven knowledge graph for food recommendation. In *SEMWEB*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890.

Abdullah Faiz Ur Rahman Khilji, Riyanka Manna, Sahinur Rahman Laskar, Partha Pakray, Dipankar Das, Sivaji Bandyopadhyay, and Alexander F. Gelbukh. 2021. CookingQA: Answering Questions and Recommending Recipes Based on Ingredients. *Arabian Journal for Science and Engineering*, pages 1–12.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sharada Prasanna Mohanty, Gaurav Singhal, Eric Antoine Scuccimarra, Djilani Kebaili, Harris Héritier, Victor Boulanger, and Marcel Salathé. 2021. The food recognition benchmark: Using deeplearning to recognize food on images. *arXiv preprint arXiv:2106.14977*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *CoCo@NIPS*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Jingxuan Tu, Eben Holderness, Marco Maru, Simone Conia, Kyeongmin Rim, Kelley Lynch, Richard Brutti, Roberto Navigli, and James Pustejovsky. 2022. Semeval-2022 task 9: R2VQ – competence-based multimodal question answering. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

A Dataset Analysis

One of the major problems we encountered during the linguistic analysis was question duplicates:

Semantically justified

- Semantic ambiguity resulting from the specific characteristics of the English language, so that it is not possible to distinguish question types by their syntactic structure or by other elements. For example, four groups of semantic roles as possible answers can appear in the same recipe
  Duplicated question: How do you mix the shrimp, pasta, butter and parsley?
  Answer 1: mix the shrimp, pasta, butter and parsley well
  Answer 2: by using a spatula

- With the same question structure, semantic ambiguity resulting from the content of the recipe, e.g. the same verb appearing twice in the text.
  Duplicated question: What should be added to the pan?
  Answer 1: the string beans and dressing
  Answer 2: the sauteed garlic, onions, ginger and string beans

In such cases it might be a better solution to list the correct answers instead of giving just one.

Semantically unjustified

- With the same question structure, referring to the same object in the recipe; these appeared mainly in the COUNTING category.
  Duplicated question: How many bowls are used?
  Answer 1: N/A
  Answer 2: 1

The aforementioned example of unjustified duplicates is associated with another problem. Namely, for many questions marked as unanswerable it was actually possible to find an answer in the recipe. We suppose that this was caused by selecting random questions from other recipes and assuming that they could not be answered based on the content of another recipe. Unfortunately, due to the relatively small variety of vocabulary related to cooking, this assumption was misleading. This can be seen especially in the categories: COUNTING, LIFESPAN and ELLIPSIS.

B Additional Experiments

B.1 Applicability to Another Domain QA

In this experiment we checked whether our system would work for other domains. We chose four instruction texts that are related to make-up techniques, furniture assembling and handmade Christmas decorations. We labeled these texts manually and asked linguists to write questions and answers bearing in mind the structure of questions and answers proposed by the organizers. They created 20 QA pairs for each text on average. The system achieved results in the range of 40%-50% EM (if we additionally included responses that are semantically correct, but not fully consistent with the answer suggested by the question authors, we reached approximately 60% EM). It is worth emphasizing that this was possible without any changes in our system.

B.2 Completeness and Correctness of Question Intents

The second experiment checked whether the questions provided by the organizers were semantically diverse and to what extent they corresponded to potential human intentions. We asked linguists to write questions related to five recipes from the validation set. Importantly, for the sake of an unbiased experiment, those were not the same people who worked on the human benchmark. The linguists engaged in this experiment had not seen questions and answers provided by the organizers, so the structure of independently written questions is not influenced by the existing dataset. They prepared about 100 question-answer pairs (20 for each recipe). After comparing the questions provided by the organizers to the ones created by our linguists, we concluded that some question types have not been included in the competition dataset:

- questions related to the amount of ingredients, e.g *How many tablespoons of vinegar should I add?*
- questions about the type of ingredients, e.g. *What kind of oil should I use for this recipe?*
- yes-no questions, e.g. *Is spinach required for this recipe?*
- questions about name or type of the dish, e.g. *What is this recipe for?*

Therefore, we have four extra categories not mentioned by task organizers. On the other hand, every

category in Section 3.1 was covered by at least one question.

It is also noteworthy that most of the questions formulated by linguists are in the first person singular instead of the second person, as the organizers propose. Also, they respond using a whole sentence rather than a single word or a short phrase. The remaining questions written by the linguists correspond to the questions categories proposed by the task organizers. This proves that the proposed Question categories are valid and reflect real human intentions. It should be emphasized that the structure of human-written questions and answers is much more varied, but they still contain keywords that can be used without problems in our question classifier.

It must be stressed that our manual annotation concerned entirely new texts, only for the purpose of these experiments. We did not use the any of the additionally annotated data to augment the datasets provided by task organizers. Therefore, the experiments did not affect out final score.

C Implementation Details

The process of searching for information in a recipe and generating answers is presented in the Algorithm 1. It utilizes information such as types of semantic labels playing the crucial role while answering a given question category. It is summarized in Table 4, which also shows regular expressions used by our classification system.

- By *event* we usually mean a verb annotated as EVENT which should match the verb from the question. If the question also includes an adverbial, it can be used to distinguish the correct *event* in the recipe.
- By *object* we mean a word or phrase, which is annotated as DROP, PATIENT or THEME and matches the *object* from the question. In some cases no *object* is provided. Then the system relies on *event* matching.
- In the COUNTING category we need to search directly for TOOLS, HABITATS or RESULTS.

If no matching *event*, *object*, HABITAT, TOOL or RESULT can be found within the recipe, the system concludes that the question is not answerable.

Example of answering can be seen in Fig 1.

Additional Remarks

To ensure higher accuracy of the results, the system has to take into account several characteristics of

Algorithm 1: Answer generation process

Input : question, recipe
Output : generated answer

question category ← predict category using regex from TABLE 4 COLUMN 2
question details ← extract details from the question (see COLUMN 3)
relevant information ← search for relevant part in the recipe using *question details*
RC threshold ← 0.98
NA threshold ← 0.99

if *relevant information* was found **then**
    answer ← generate answer for given *question category* according to COLUMN 4
else
    answer ← use answer predicted by Electra Extractive QA[1]
    if *confidence* < *RC threshold* **then**
        answer ← N/A

if *N/A Classifier*[1,2] *output* = *N/A* and *confidence* >= *NA threshold* **then**
    answer ← N/A

return *answer*

[1] Electra Extractive QA and N/A Classifier are used only for some categories

[2] N/A Classifier was added after competition end

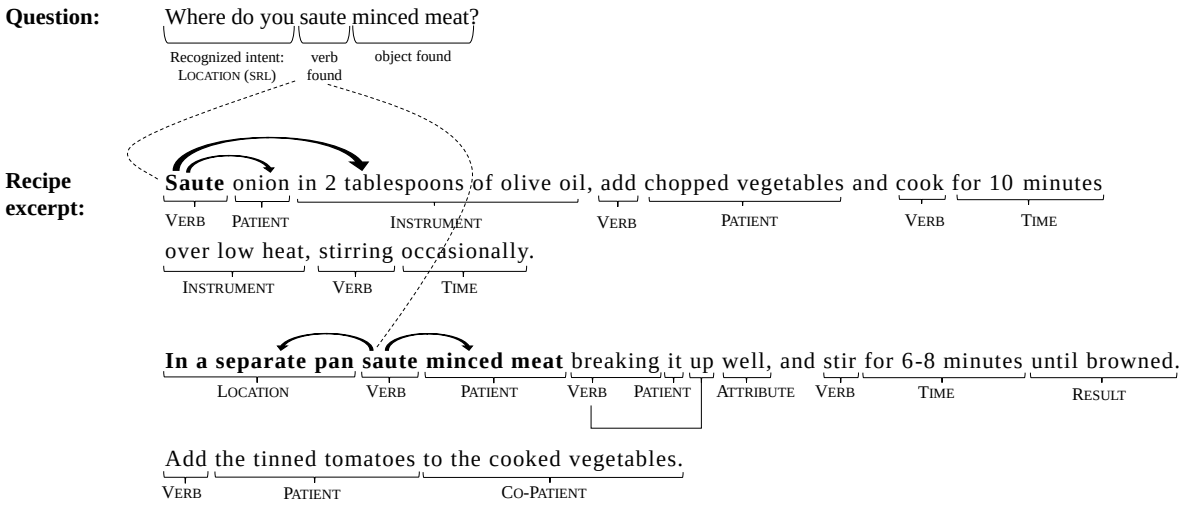
Table 4: Summary of question handling. Columns left-to-right: category, regex used for initial classification, semantic information used to search for the answer in the recipe, information used to generate the final response.

Category	Regex Pattern	Searched Label	Answer Generation
COUNTING TIMES	How many times	tools or habitats	count found occurrences
COUNTING ACTIONS	How many actions	result and corresponding event	count found occurrences
COUNTING USES	How many .* are used	tools or habitats	count found occurrences
ELLIPSIS	What should	event and (tool or habitat)	drops, ingredients
LOCATION (CRL)	Where should you	event and object	habitat
LOCATION (SRL)	Where do you	event and object	location, destination, co-patient or co-theme
METHOD	How do you	event and (object or ingredients)	verb, object, one of: tool, instrument, attribute, goal
LIFESPAN (HOW)	How did you get	result and corresponding event	verb, drops, patients, tools, habitats
LIFESPAN (WHAT)	What's in	result and corresponding event	ingredients (if patient or theme), drops
EVENT ORDERING	.* which comes first	both events	use the preceeding one
RESULT	To what extent	event and object	result
TIME	For how long	event and (object, attribute or purpose)	time
EXTENT	By how much	event and object	extent
PURPOSE	Why do you	event and object	cause or purpose
CO-PATIENT	What do you .* with	event and object	co-patient or co-theme
SOURCE	From where	event and object	source
LOCATION CHANGE	Where was .* before	event and object, and all previous events for the same object	previous habitat different from the one in the starting event

the R2VQ dataset:

- SRLs are represented as columns, within which objects are connected to the verb. Each subsequent verb within a sentence has its own column with corresponding objects. Iterating over each column separately appears to be very helpful in terms of associating verbs with proper objects.
- Each SRL starts with the head (the label starts with the letter B). If the phrase contains multiple words, the head is followed by the body (the label starts with the letter I or D). We found that concatenation of the full-length expression (using B and I as indicators) improves the quality of the identification process.
- Tokens whose CRL is TOOL, HABITAT, EXPLICITINGREDIENT or IMPLICITINGREDIENT are supplied with the index of the verb to which they relate. In some cases, using that information is extremely helpful as it allows for unambiguous identification of the relationship between the verb, the object and the answer.
- While an object in the question is created using a HIDDEN ROLE, it is needed to singularize each part of it. For example, if Drop="**limes**.3.1.9:**ginger**.3.1.1:**onions**.2.1.7" there is a great chance that it will appear in the question in the form of *the lime, ginger and onion*. On the contrary, when CRL or SRL were used to create the question, they will most likely appear as an unchanged span from the passage.

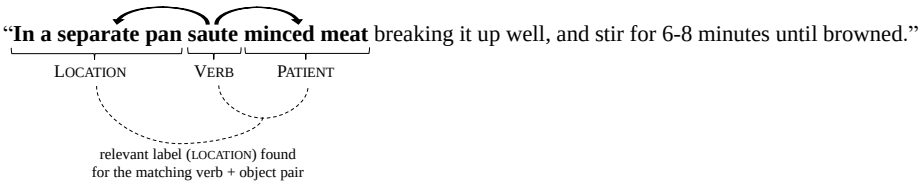
Figure 1: Example of the sematic-role-based answer generation.



Procedure:

1. Recognized intent = LOCATION (SRL)
   Verb = saute
   Object = minced meat

2. Relevant Sentence (VERB = **saute** & PATIENT = **minced meat**):
   There are two sentences with verb *saute*.
   The model chooses the one whose object (PATIENT) is *minced meat*.



Answer: in a separate pan