BioMedLAT Corpus: Annotation of the Lexical Answer Type for Biomedical Questions

Mariana Neves and Milena Kraus
Hasso-Plattner Institute
August-Bebel-Str. 88
Potsdam, 14482, Germany
mariana.neves@hpi.de, milena.kraus@hpi.de

Abstract

Question answering (QA) systems need to provide exact answers for the questions that are posed to the system. However, this can only be achieved through a precise processing of the question. During this procedure, one important step is the detection of the expected type of answer that the system should provide by extracting the headword of the questions and identifying its semantic type. We have annotated the headword and assigned UMLS semantic types to 643 factoid/list questions from the BioASQ training data. We present statistics on the corpus and a preliminary evaluation in baseline experiments. We also discuss the challenges on both the manual annotation and the automatic detection of the headwords and the semantic types. We believe that this is a valuable resource for both training and evaluation of biomedical QA systems. The corpus is available at: https://github.com/mariananeves/BioMedLAT.

1 Introduction

Question answering (QA) systems are able of providing exact answers for input questions (Athenikos and Han, 2010; Neves and Leser, 2015). However, coherent answers can only be returned if the system correctly understands the question that is posed. In a QA system, the question processing (or understanding) step includes many components, such as linguistic analysis (e.g., tokenization, part-of-speech tagging, semantic role labeling and parsing), question type identification (e.g., yes/no, factoid, definition), lexical answer type (LAT) identification (e.g., protein or disease name) and query construction.

In this work we focus on the LAT component of a QA system, i.e., the identification of the expected type of the answer that needs to be returned. This is especially important for factoid questions, i.e., questions that expect an exact and short answer in return, such as a protein or disease name. The LAT task can be divided into two steps: (i) recognition of the headword, followed by (ii) its classification into predefined type(s). For instance, in the question *"What hand deformities do patients with Apert syndrome present with?"*, "deformities" is the headword of the question while "Sign or Symptom" is a possible expected type.

Although the field of question answering for biomedicine has evolved in the last years thanks to the many editions of the BioASQ challenges (Tsatsaronis et al., 2015), researchers still miss important resources to support both development and evaluation of biomedical QA systems. BioASQ has provided the community with the most important benchmark in this domain but the dataset does not include information on the expected LAT's. The latter is an important detail, which enables both the evaluation of the LAT identification component in biomedical QA systems as well for training in machine-learning-based methods.

We manually annotated a set of 643 questions from the BioASQ training data with the headword and the corresponding UMLS semantic type. We evaluated our annotations using a baseline approach based on dictionary-based matching of UMLS-derived dictionaries. In this paper, we describe the guidelines

Proceedings of the Open Knowledge Base and Question Answering (OKBQA) Workshop,
pages 49–58, Osaka, Japan, December 11 2016.

and also the results of our annotation process. We evaluate and discuss on the statistics of the annotations, on the complexity of the annotation task and on the error analysis of our baseline approach.

2 Related Work

Construction of a classification of types is common in other domains, such as the PICO framework in the medical domain (Armstrong, 1999). A good overview of taxonomies for the medical QA is provided in (Athenikos and Han, 2010). The UMLS semantic types have also been successfully used for the medical domain, such as in (Kobayashi and Shyu, 2006). During the development of the INDOC question answering system (Sondhi et al., 2007), the authors carried out an analysis of the frequency of the UMLS semantic groups in 106 questions from the OHSUMED collections. The main objective of the analysis was to define different weights for each semantic group in the INDOC system. They report that the most frequent types where the following: "Concepts & Ideas" (CONC), "Disorders" (DISO) and "Procedures" (PROC).

Since the start of the BioASQ challenges, which promoted many innovations in biomedcial QA, some participants have also tried to predict semantic types for factoid questions, as described in details below. However, we are not aware of any previous publications on manual annotation of semantic types and headwords for the BioASQ dataset.

One of the first works to identify the LAT for the BioASQ dataset was carried out in (Weissenborn et al., 2013). Their system classified the question into three classes: (1) What/Which questions, (2) Where-questions, and (3) decision questions. They relied on regular expressions to extract the headword of the question, but they did not attempt to predict the expected types of the answer. They relied on Metamap for mapping the headword to one of the UMLS semantic types (Aronson and Lang, 2010).

In (Yang et al., 2015), the authors extended the work of (Weissenborn et al., 2013) and considered two more classes: "choice" and "quantity". The recognition of the concepts in the question was also performed using Metamap but variants were added with the UMLS Terminology Services (UTS)[1].

The system developed by the Fudan team (Peng et al., 2015) automatically classified the questions into some few semantic types, namely: (1) disease, (2) drug, (3) gene/protein, (4) mutation, (5) number and (6) choice. The sixth type did not indicate a specific semantic type and it was used in situations in which the possible answers are provided in the question, The system used some rules to identify the expected types. The semantic types were also used by the PubTator tool for named-entity recognition. The extracted entities were the candidate exact answers for the question.

The work of (Yenala et al., 2015) was restricted to the identification of the headwords and they developed an algorithm for the so-called "Domain Word Identification". However, they did not attempt to identify the semantic type of the extracted domain words. Instead, the headword is used to filter out words which are not relevant for the passage retrieval step while the extraction of the exact answer was only based on linguistic features and text similarity.

Finally, in the YodaQA system (Baudis and Sediv, 2015), the headword of the question was extracted and its LAT was identified using the titles of the documents in Wikipedia, i.e., by relying on Wikipedia's classes. As an extension of the system to the biomedical domain, they also considered the Gene Ontology (GO) using the GOLR endpoint by considering the type field as the LAT of the question.

3 Corpus Annotation

In this section, we describe our resources, the annotation process and our annotation guidelines for headwords and semantic types.

3.1 Data

We relied on two main resources to perform the annotation of the headwords and the assignment of the semantic types: the BioASQ datasets of questions and the UMLS semantic types.

[1] https://uts.nlm.nih.gov/home.html

BioASQ questions. We utilized the questions made available during the first, second and third editions of the BioASQ challenge[2] (Tsatsaronis et al., 2015). The BioASQ challenge includes four types of questions, namely "yes/no", "summary", "factoid" and "list". The "yes/no" question requires either "yes" or "no" answer, while the "summary" question expects a short paragraph as answer. Neither of them require the identification of the semantic type of the answer. Therefore, we carried out manual annotations only for the "factoid" and "list" questions, which expect one or more exact answer(s) of a certain semantic type in return. We downloaded the current BioASQ training dataset[3] in the JSON format and extracted the following information for the "list" and "factoid" questions: (i) question identifier (tag "id"), (ii) question text (tag "body"), and (iii) exact answers (tag "exact_answer").

UMLS semantic types. The UMLS semantic types[4] are a set of categories (and groups of categories) that are used to cluster concepts of the same type in the Metathesaurus (Bodenreider, 2004). It currently contains 133 types divided into 14 groups. We find it is an appropriate resource for the annotation of our corpus given the amount of research that makes use of the UMLS database and the Metamap tool (Aronson and Lang, 2010). For instance, the UMLS semantic types were integrated into the BioTop ontology (Schulz et al., 2009) and previously used for medical QA (Kobayashi and Shyu, 2006). We downloaded the list of semantic types in the plain text format[5] and used it for our annotation.

3.2 Manual Annotation Process

We performed the annotation on the brat annotation tool (Stenetorp et al., 2012). We created the document files by concatenating the text of the question and the exact answers from the BioASQ gold standard (GS) file. We included the exact answer(s) to support the manual assignment of the semantic type (as discussed in the guidelines below). Figure 1 shows an example of an annotated question in brat.

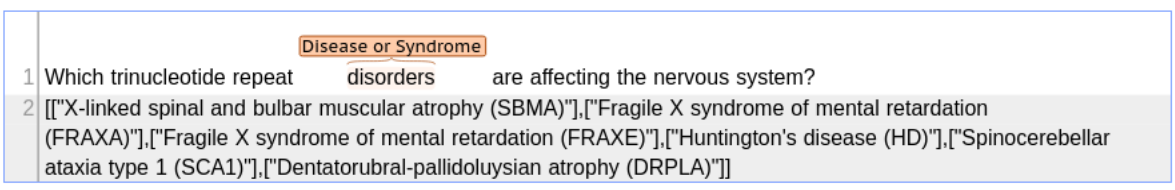

Figure 1: Screen-shot of annotation in brat annotation tool. We included both the question (line 1) and the answers (line 2), just as provided in the BioASQ training set.

Two annotators conducted the manual annotation process: one is a PhD student in computer science who has majored in biotechnology (genetics, biochemistry and bioinformatics) and the other is a computer scientist with deep knowledge and ten years of experience on biomedical natural language processing. Each annotator performed the annotations and then a final version of the corpus was created during many consensus sessions, in which notes were taken on disagreements on both the semantic types and groups.

3.3 Annotation Guidelines

We defined guidelines for the annotation of both the headword of the question and the assignment of its semantic type.

3.3.1 Headwords

We define headword as the minimum text span that identifies the expected LAT. Therefore, it is not limited to the words following the Wh- question word. More details are presented below:

[2] http://bioasq.org/
[3] http://participants-area.bioasq.org/Tasks/4b/trainingDataset/
[4] https://semanticnetwork.nlm.nih.gov/
[5] https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt

1. The text span of the headword should include enough words to support the identification of its semantic type. For instance, in the question "Which are the synonyms of prostate-specific antigen?" (id 5171651e8ed59a060a000009), the headword "synonym" is not meaningful enough to support the assignment of the semantic type, while the phrase "synonyms of prostate-specific antigen" indicates that the answer should be an antigen.

2. The headword should not include unnecessary words that qualify the headword but that have no influence on the decision of the semantic type, such as "of prostate specific" in the previous example. In this case, the headword was restricted to "synonyms of antigens" (discontinuous annotation).

3. In the case of choice questions, multiple headwords were annotated. For instance, the question "Is cancer related to global DNA hypo or hypermethylation?" (id 516e5f10298dcd4e5100007c) has two headwords ("global DNA hypo" and "hypermethylation").

4. Some questions have no explicit headword, i.e, the type of the target is given by the Wh- particle and by the words of the question. For instance, the question "Where is X-ray free electron laser used?" (id 51475d5cd24251bc0500001b) requires a location as answer, given by the "where" particle. However, this particle can lead to different UMLS semantic types depending on the context. For instance, in the question "Where in the cell do we find the protein Cep135?" (id 51596a8ad24251bc0500009e), the answer is a cell component, , i.e., UMLS semantic type "T026". On the other hand, "centromeres" is the answer to the question "Where is the histone variant CENPA preferentially localized?" (id 52fe52702059c6d71c000078), thus, a nucleotide sequence (T086).

3.3.2 Semantic types

We assigned one or more semantic types to the identified headword. More details on the annotation are presented below:

1. The semantic types should be defined not only based on the headword, but also on the exact answers included in the gold standard dataset. For instance, for the question "Which are the best treatment options to treat Helicobacter pylori?" (id 518cb5ab310faafe08000008), the system could return either clinical drugs or procedures as answer. However, given that the gold standard includes only clinical drugs in the exact answer, e.g. "amoxicillin" and "metronidazole", we mapped the headword "treatment" to the clinical drug type.

2. In cases in which the question is composed of more than one sentence, the decision should take into account the complete text and not only the question, as in the following example: "A common problem in proteomics is the contamination of samples with exogenous proteins (often from other species). These proteins can be found in specific databases. List some contaminants." (id 515d7693298dcd4e5100000c). It consists of multiple sentences that are descriptive of the required semantic type. While "contaminants" as headword extracted only from the last phrase would include many possible semantic types, such as the complete group of chemicals or some types of the group organisms, the headword "protein" found in the previous sentences specify the semantic type to be "Amino Acid, Peptide or Protein".

3. We assigned one or multiple semantic types if the answer contained multiple, different types. For example, the answers to "Which substances are dangerous to g6PD deficient individuals?" (id 5314b20bdae131f847000005) are "fava beans" and "primarquine" amongst others. While beans belong to the type "Objects - Food", primarquine can be categorized as "Chemical - Clinical Drug". There were only a couple of such cases.

4 Experiments

In this section, we describe a simple baseline experiment that we performed for evaluation of our corpus. It included both the extraction of the headword and the identification of the LAT. Similar to previous

works, we extracted the headword based on both NER and simple heuristics. We used the following regular expression to process a question and to extract its headword: ((what |where |which |who) (<(plural) noun> is| are .*))

After the headword extraction, we performed an NER step on the question. We matched words in the question to UMLS concepts based on various UMLS ontologies. Given the concepts identified in the question, we checked their overlap with the previously identified headword.

For instance, for the question "Which genes have been proposed as potential candidates for gene therapy of heart failure?", we identified "genes" as the headword, using the above regular expression. The same word "genes" also matched the UMLS concept "C0017337" in the NER step. Finally, as the concept "C0017337" is linked to the type "Gene or Genome" (T028), this is the LAT of the question.

5 Results

In this section, we present the details of our corpus and results from our baseline experiments.

5.1 Statistics of the Annotations

The BioASQ training data (cf. 3.1) contains a total of 654 question annotated as "factoid" or "list". We assigned one or more semantic types for a total of 643 questions, as we removed eight BioASQ questions that we found were incorrectly classified as factoid/list (cf. 6.1). We created 647 annotations with a total of 53 distinct semantic types (from 133 UMLS semantic types) and 343 distinct headwords.

Table 1 displays a list of the top eight semantic types that each occurred more than 20 times in our corpus. The number of annotations of these top eight semantic types add up to 406, which corresponds to around 63% of the whole data set. Thus, 45 types account for the other 37% of the annotations.

No. Annotations	Semantic Group	Semantic Type
115 (17.8%)	Chemicals & Drugs	Amino Acid, Peptide, or Protein
72 (11.1%)	Disorders	Disease or Syndrome
62 (9.6%)	Genes & Molecular Sequences	Gene or Genome
40 (6.2%)	Disorders	Sign or Symptom
34 (5.3%)	Chemicals & Drugs	Enzyme
32 (4.9%)	Chemicals & Drugs	Clinical Drug
27 (4.2%)	Physiology	Genetic Function
24 (3.7%)	Genes & Molecular Sequences	Nucleotide Sequence

Table 1: List of the eight top semantic types which occur more than 20 times in the corpus.

Alternatively, QA systems could also consider our annotations only on the level of semantic groups. The 53 annotated semantic types correspond to 11 of the 15 UMLS semantic groups. Table 2 shows the distribution of our annotations over the various semantic groups.

No. Annotations	Semantic Group	No. Annotations	Semantic Group
218	Chemicals & Drugs	24	Phenomena
117	Disorders	21	Anatomy
88	Genes & Molecular Sequences	18	Objects
61	Concepts & Ideas	14	Living Beings
46	Procedures	2	Activities & Behaviors
38	Physiology		

Table 2: List of the eleven semantic groups included in the corpus.

We annotated 343 distinct headwords. The most frequent headwords, i.e., the ones which occur at least ten times in the corpus, are the following: genes (26), proteins (21), protein (19), gene (16), disease (13), How many (11), drugs (10) and diseases (10).

In Table 3, we list the most ambiguous headword, i.e., headwords that can refer to more than one semantic type. This situation was prevalent even for headwords which seem unambiguous at first glance, such as ”gene” and ”protein”. Some headwords, such as ”treatment”, were ambiguous even with respect to the group, as clinical drugs and therapeutic procedures belong to different semantic groups. This was also the case of the ”methods” headword which may also refer to a tool name, thus the semantic type ”Manufactured Object”.

Headword	**Semantic Types**
genes	”Gene or Genome”, ”Classification”, ”Amino Acid, Peptide, or Protein”
treatment	”Therapeutic or Preventive Procedure”, ”Clinical Drug”
methods	”Molecular Biology Research Technique”, ”Research Activity”, ”Manufactured Object”
drugs	”Clinical Drug”, ”Chemical”
inhibitors	”Organic Chemical”, ”Clinical Drug”, ”Chemical”
mutations	”Genetic Function”, ”Gene or Genome”, ”Amino Acid, Peptide, or Protein”
factors	”Amino Acid, Peptide, or Protein”, ”Disease or Syndrome”, ”Conceptual Entity”

Table 3: List of some of the ambiguous headwords in the corpus.

On the other hand, very few semantic types were clearly not ambiguous in our corpus, such as the following ones: ”Body Location or Region” (headword ”region”) and ”Virus” (headwords ”virus” and ”viruses”). Although some other semantic types have only one headword in our corpus, these are clearly not the only headwords with which we could refer to the type, but rather that these types are rare in the corpus. Examples of such types are the following: ”Group” from ”Living Beings” (headword ”kingdom”), ”Inorganic Chemical” (head word ”deficiency”) and ”Intellectual Product” (headword ”articles”). The most ambiguous type is ”Amino Acid, Peptide, or Protein” with a total of 55 headwords. Some more examples of very ambiguous semantic types and the corresponding headwords are shown in Table 4.

Semantic Type	**Headwords**
Cell Component	localization, organelles, cytoplasmic nuclear, structures, subcellular localization, Where in the cell, Where localized
Manufactured Object	software tools, database, databases, bioinformatics tools, biomedical text mining tools, tools, programs, systems, methods, computer programs, content, computational tools
Gene or genome	genes, variant, chromosomes, polymorphisms, orthologs, gene, classes, mutations, genetic determinant, members/isoforms, oncogenes, target, genetic basis, Genes, mutation, gene(s), gene chromosome

Table 4: List of ambiguous semantic types and their respective headwords.

5.2 Evaluation of the Experiments

From a total of 643 questions, our baseline experiment correctly detected the semantic types for 184 (28.6%) questions and the semantic groups for 395 (61.4%) of the questions. The most frequent semantic types that were correctly detected were the following: ”Amino Acid, Peptide, or Protein” (58), ”Gene or Genome” (T028) and ”Disease or Syndrome” (27). These are also the most frequently annotated types in the corpus, as presented in Table 1. Consequently, the most frequent groups correctly detected by our system were the following: ”Chemicals & Drugs” (212), ”Disorders” (54) and ”Concepts & Ideas” (47).

We could not correctly detect many of the semantic types in our corpus. Table 5 summarizes our most frequent errors. All of our top errors are failures to detect the ”Amino Acid, Peptide, or Protein” types, given that it contains a variety of headwords. Finally, many semantic groups that we failed to detect were from the very abstract category ”Concepts & Ideas”.

No. errors	Correct semantic type	Detected semantic type
50	Amino Acid, Peptide, or Protein	Biologically Active Substance
27	Amino Acid, Peptide, or Protein	Quantitative Concept
20	Amino Acid, Peptide, or Protein	Intellectual Product
20	Amino Acid, Peptide, or Protein	Cell Component
19	Amino Acid, Peptide, or Protein	Element, Ion, or Isotope
19	Amino Acid, Peptide, or Protein	Spatial Concept
No. errors	**Correct semantic group**	**Detected semantic group**
116	Concepts & Ideas	Chemicals & Drugs
66	Concepts & Ideas	Disorders
52	Anatomy	Chemicals & Drugs
29	Concepts & Ideas	Procedures
28	Chemicals & Drugs	Concepts & Ideas

Table 5: List of the most frequent errors for the detection of semantic types and groups.

6 Discussion

In this section, we discuss some of the challenges we encountered during the annotation of the questions as well as the results we obtained with our approach.

6.1 Challenges in the Annotation Task

We faced many challenges while manually annotating the headwords and the semantic types in the BioASQ questions. These issues range from questions that might have been mistakenly classified as ”factoid” to questions, answers which were too abstract and semantic types which were difficult to identify.

Non-factoid questions. We came across some questions in BioASQ that were probably mistakenly annotated as ”factoid” or ”list”, when they should have been classified as ”summary” instead. For instance, the question ”Why is lock mass used in Orbitrap measurements?” (id 530b01a6970c65fa6b000008) clearly expects more than one short answer in return, given the ”why” particle, and indeed has the following sentence as exact answer:”The lock mass is a compound of known mass and is used to compensate for drifts in instrument calibration.” We also found some ”yes/no” questions among the list of questions that we analyzed, such as ”Is there a crystal structure of the full-length of the flaviviridae NS5(Methyltransferase - RNA depended RNA Polymerase)?” (id 532aad53d6d3ac6a34000010), to which the name of the crystal structure was annotated as answer, though. Furthermore, ”Is there a crystal structure of Greek Goat Encephalitis?” (id 532819afd6d3ac6a3400000f), whose answer ”No crystal structure of Greek Goat Encephalitis found” is clearly equivalent to a ”no” answer. In summary, we removed the following eight questions from our corpus: 54fc4e2e6ea36a810c000003, 530b01a6970c65fa6b000008, 530cf54dab4de4de0c000009, 531b2fc3b166e2b80600003c, 530cf4e0c8a0b4a00c000002, 5348307daeec6fbd07000011, 532819afd6d3ac6a3400000f, 532aad53d6d3ac6a34000010.

Errors in the question formulation. We believe that we found some errors in the question formulation in a way that it leads to wrong semantic types and headwords. For instance, we expected a function as answer to the question ”Which hormone receptor function is altered in patients with Donohue syndrome?” (id 2b4/5314bd7ddae131f847000006). However, ”insulin”, i.e., a hormone, is the answer instead. Therefore, we believe the question should be rephrased to, e.g., ”For what hormone is the receptor function altered in patients with Donohue syndrome?”. Two other examples of this situation are the following questions: ”Which hormone deficiency is implicated in the Costello syndrome?” (id 53130a77e3eabad02100000f) and ”Which hormone abnormalities are characteristic to Pendred syndrome?” (id 53148a07dae131f847000002). Curiously, all examples expect a hormone name

as answer. In one particular case, we expected a number to be the answer, but the BioASQ gold standard returns a list of cancer types: "How many different subtypes of thyroid cancer exist?" (id 5503145ee9bde69634000022). We did not change the original questions during our annotation.

Challenges on the headwords. For some questions, no headword was explicit and we had to highlight the text span that gave some hints on the headword instead. The question "What is SCENAR therapy used for?" (id 535d69177d100faa09000003) is a good example. It expects disease names as answers and we chose to highlight the discontinuous annotation "what...used for" as headword. A similar example is shown in the question "What does mTOR stands for?" (id 5505a587f73303d458000005), for which we annotated the headword "what...stands for".

Challenging answers QA is a challenging task in itself, but we found questions which were particularly challenging with regard to assigning the semantic type and also for getting the expected answer. For some questions, many other words needed to be taken into account in order to identify the LAT. For instance, the question "What is being measured with an accelerometer in back pain patients" (id 533f9df0c45e133714000016) has the following answers: "Physical activity", "Constant Strain Postures", "Standing time", "Lying time". This is a rather abstract question with answers which do not easily fit any of the UMLS types. We decided to categorize the answers as "Conceptual Entity".

One particular question in the dataset includes two questions with two distinct semantic types: "How many and which are the different isoforms for the ryanodine receptor?" (id 3b1/54db7217c4c6ce8e1d000003). This is indeed a question that a user could ask and the system should preferably provide not only the list of isoforms but also the total number of them. BioASQ provides only the first of the answers but we annotated two headwords ("how many" and "different isoforms") and assigned the two corresponding semantic types, i.e., "Quantitative Concept" and "Chemicals & Drugs-Receptor".

6.2 Agreement on the Annotations

We computed a total of 66 (10.2%) disagreements on the group-level and 49 (7.6%) disagreements on the type-level. This is not surprising, given the challenges discussed above. In general, disagreement on document-level were related to choosing either the "Phenomena" (PHEN) or the "Physiology" (PHYS) groups. Disagreements on the type-level were also frequently related to different types of the "Chemicals & Drugs" (CHEM) and "Genes & Molecular Sequences" (GENE). One example of a divergence on the type can be found in the question "Which are the DNA (cytosine-5-)-methyltransferases inhibitors?" (id 5165932e298dcd4e51000059). One of the annotators assigned the more general "Chemical" types while the other one assigned the "Organic chemical" type. As both types are correct we decided for the more precise annotation "Organic Chemical".

Disagreements on group level also occurred on mistakes of one of the annotators when assigning the "Gene or Genome" (T028) type (group GENE) when a protein (type T116 of group CHEM) was expected. An accurate discrimination of genes, any types of intermediate RNA and the resulting proteins is inherently complex and may be even impossible. This can be exemplified by the question "What are the major classes of retrotransposons active in the human genome?" (id 517843638ed59a060a000036). One annotator assigned the type "Gene or Genome" whereas the term gene can be misleading as retrotransposons can contain no gene-like information (e.g. the Alu element) or multiple genes in one transposon (e.g. LTR retrotransposons). The other annotator assigned a type from the "Classification" group, which is a more general annotation.

6.3 Quality of the Annotations

As discussed above, annotating the headwords and the semantic types is a complex and subjective task. We checked the gold-standard answers from BioASQ upon deciding the semantic types and the two annotators achieved a good agreement score for the group level. However, we neither retrieved nor checked whether the answers have a corresponding concept in UMLS.

Furthermore, most annotations are represented by just a few semantic types. A second iteration of annotation might result in a better distribution of types of the same group. This might be the case espe-

cially in the "Disorders" group where most annotations were concentrated on the "Disease or Syndrome" type. Finally, four semantic groups were not annotated in our corpus:"Devices", "Geographic Areas", "Occupations" and "Organizations". Although we might have missed some of these groups during our annotation, our annotations could also serve as feedback for the BioASQ organizers on new topics to address for the next editions of the challenge.

7 Conclusions

We presented our annotation of the BioASQ dataset of biomedical question with respect to headwords and the expected lexical answer types. We manually annotated a set of 643 questions and we provided an overview on the annotations, disagreements and possible mistakes in the questions. We also presented a comprehensive discussion on the challenges that we faced during the annotation process, which could also be translated to challenges to the question answering systems. Finally, we ran baseline experiments to evaluate the extraction of headwords and semantic types.

Acknowledgements

We would like to thanks support from the students Maximilian Götz, Marcel Jankrift, Julian Niedermeier, Toni Stachewicz and Sören Tietböhl.

References

[Armstrong1999] E. C. Armstrong. 1999. The well-built clinical question: the key to finding the best evidence efficiently. *WMJ*, 98(2):25–28.

[Aronson and Lang2010] Alan R Aronson and Franois-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

[Athenikos and Han2010] Sofia J. Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1 – 24.

[Baudis and Sediv2015] Petr Baudis and Jan Sediv. 2015. Biomedical question answering using the yodaqa system: Prototype notes. In Linda Cappellato, Nicola Ferro, Gareth J. F. Jones, and Eric SanJuan, editors, *CLEF (Working Notes)*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[Bodenreider2004] Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–D270, Jan.

[Kobayashi and Shyu2006] Tetsuya Kobayashi and Chi-Ren Shyu. 2006. Representing clinical questions by semantic type for better classification. *AMIA Annual Symposium Proceedings*, 2006:987–987.

[Neves and Leser2015] Mariana Neves and Ulf Leser. 2015. Question answering for biology. *Methods*, 74:36 – 46.

[Peng et al.2015] Shengwen Peng, Ronghui You, Zhikai Xie, Beichen Wang, Yanchun Zhang, and Shanfeng Zhu. 2015. The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering. In *Working Notes for CLEF 2015 Conference, Sheffield, UK, September 15-18, 2014.*, pages 1337–1347.

[Schulz et al.2009] S. Schulz, Elena Beisswanger, Lszl van den Hoek, Olivier Bodenreider, and Erik van Mulligen. 2009. Alignment of the umls semantic network with biotop: Methodology and assessment. *Bioinformatics*, 25(12), June.

[Sondhi et al.2007] Parikshit Sondhi, Purushottam Raj, V. Vinod Kumar, and Ankush Mittal. 2007. Question processing and clustering in indoc: A biomedical question answering system. *EURASIP J. Bioinformatics Syst. Biol.*, 2007:1:1–1:7, July.

[Stenetorp et al.2012] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

[Tsatsaronis et al.2015] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

[Weissenborn et al.2013] Dirk Weissenborn, George Tsatsaronis, and Michael Schroeder. 2013. Answering factoid questions in the biomedical domain. In Axel-Cyrille Ngonga Ngomo and George Paliouras, editors, *Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering, Conference and Labs of the Evaluation Forum 2013 (CLEF 2013)*.

[Yang et al.2015] Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, , and Eric Nyberg. 2015. Learning to answer biomedical factoid & list questions: Oaqa at bioasq 3b. In *Working Notes for CLEF 2015 Conference, Sheffield, UK, September 15-18, 2014.*

[Yenala et al.2015] Harish Yenala, Avinash Kamineni, Manish Shrivastava, and Manoj Chinnakotla. 2015. Iiith at bioasq challange 2015 task 3b: Bio-medical question answering system. In *Working Notes for CLEF 2015 Conference, Sheffield, UK, September 15-18, 2014.*