Annotating Question Types in Social Q&A Sites

Kateryna IGNATOVA, Cigdem TOPRAK [1], Delphine BERNHARD and
Iryna GUREVYCH
Ubiquitous Knowledge Processing (UKP) Lab, Computer Science Department
Technische Universität Darmstadt, Germany
www.ukp.tu-darmstadt.de

Abstract. In all domains, including eHumanities, it is crucial to understand how people seek information and what kinds of questions they ask. In this paper, we present an annotation study of domain-specific questions collected from the current leading social Question and Answer site, namely Yahoo! Answers. We define an annotation scheme with 9 question types and additional attributes to identify unclear and opinion questions. We show that annotating questions extracted from social media content is a difficult task due to errors and ambiguities in question formulations. However, we obtain good to very good inter-annotator agreement on all but one of the defined question types.

Keywords. question type, annotation study, social Q&A

1. Introduction

Information search and the ways users express their information needs in the form of questions is a fundamental issue in all domains, including eHumanities. Indeed, information search is an important tool for eHumanities researchers looking for information in their discipline-specific information repositories. It is therefore crucial to understand how people ask questions when seeking information on a given topic.

This is also highly relevant for automatic Question Answering (QA) systems. Typical QA systems rely on a question type classification which circumscribes the kind of questions that the system is able to answer. Most of the existing open-domain QA systems utilize question type classification schemes tuned to answer a restricted set of factoid, definition or list questions from the TREC (Text REtrieval Conference) or CLEF (Cross Language Evaluation Forum) QA evaluation campaigns [1]. QA systems aiming to cope with more complex user questions necessitate broader question type classifications, based on real user questions collected in an authentic setting.

In this article, we propose to use the wealth of questions available in the Yahoo! Answers (YA) social Question and Answer (Q&A) site[2] to perform a thorough study of

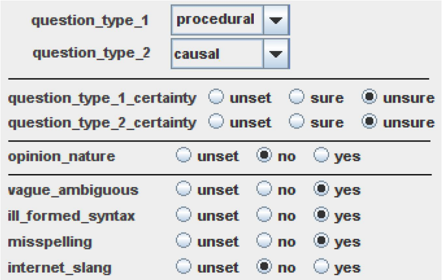
[1] Corresponding Author: Cigdem Toprak, UKP Lab, Technical University of Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany; E-mail: c_toprak@tk.informatik.tu-darmstadt.de

[2] According to [2], Yahoo! Answers is the current leader in the social Q&A market

the types of questions users ask online, getting closer to realistic use cases for automatic QA systems. To this aim, we introduce a question type annotation scheme, enhanced with attributes to identify unclear and opinion questions, aiming at answering the following research questions: (i) what types of information seeking questions are actually asked in social Q&A sites, (ii) what is the proportion of domain-specific questions that entail opinionated answers, and (iii) how well written are questions asked on social Q&A sites.

2. Annotation Scheme

The goal of our annotation study is to capture several kinds of information about user questions, as shown in Figure 1.[3]



question_type_1 procedural
question_type_2 causal

question_type_1_certainty ○ unset ○ sure ● unsure
question_type_2_certainty ○ unset ○ sure ● unsure

opinion_nature ○ unset ● no ○ yes

vague_ambiguous ○ unset ○ no ● yes
ill_formed_syntax ○ unset ○ no ● yes
misspelling ○ unset ○ no ● yes
internet_slang ○ unset ● no ○ yes

Figure 1. Example annotation of the question: *in Excel, regression problem, "Input data conatins non-numeric data"?*

As a basis for question type annotation we used the scheme developed by [3]. Our slightly modified scheme is detailed in Table 1. To account for ambiguous and multiple sentence questions, we allow the annotators to assign two question type labels to each question.

A better understanding of opinion questions is required by multi-perspective QA systems [4]. For this reason, we additionally define the binary opinion attribute to assess the user's request for opinions or suggestions on the question's topic (e.g., *what can we do for quality and improvement in higher or lower education?*)

To identify questions problematic for manual and automatic analysis, we use additional attributes to assess clarity on the semantic (*ambiguity*), syntactic (*ill-formedness*), and lexical (*slang*, *misspellings*) levels.

3. Annotation Study

The follow-up experiment involved three annotators: two students of English linguistics (later, A1 and A2) and a co-author of the paper (A3). The annotation was performed using MMAX2 [5].

[3] We deliberately kept spelling errors in the examples.

Proposed Type	**Examples from Yahoo! Answers**
Concept Completion	*which r websites to learn web design?*
Definition	*what is KPO (knowledge processing and out sourcing)?*
Procedural	*How do I create a risk management database?*
Comparison	*what is the difference between retesting and regression testing?*
Disjunctive	*Which one is better to use for speech recognition and image processing..C,C++,VC++ or Matlab?*
Verification	*Does the linear regression of a data set pass through the centroid of the data set?*
Quantification	*how many bytes of storage are available just using the 6800's data registers?*
Causal	*why is 0.05 used as s significant value in data analysis?*
General Information Need	*i have a hard time dealing with database management can anyone please help me?*

Table 1. The proposed question type classification scheme based on [3].

3.1. Experimental Setup

We compiled a dataset by extracting questions from the YA website using the Yahoo! Answers API[4] focusing on the domains of Data Mining, Natural Language Processing (NLP), and eLearning. The resulting dataset contained 805 questions, 50 of which appeared twice since they occurred in several categories. We did not exclude repeated questions to use them later for assessing intra-annotator agreement. We divided the dataset of 805 questions into 50 training questions and 755 questions for the annotation study.

To measure inter- and intra-annotator agreement, we use the Kappa statistic[5] [7] and, basing upon the ideas presented in [8], define two basic ways to assess agreement. **Partial Overlap (PO)** requires the agreement of the annotators on *at least one* label, i.e. partial agreement is counted as agreement. **Complete Overlap (CO)** requires the agreement on *both* labels. Furthermore, we are interested in how well the annotators agree on the question type in those cases when they are confident about their choices. Thus, we additionally calculate agreement when the *certainty* attribute is labeled as "sure" ($\textbf{PO}_{sure}$ and $\textbf{CO}_{sure}$).

3.2. Experimental Results

Table 2 displays the distribution and the distinguishability of the question types[6]. Almost half of the questions (46.3%) were classified as *Concept Completion*. Such a large proportion shows that it might be relevant to refine the *Concept Completion* type in fu-

[4] `http://developer.yahoo.com/answers/`

[5] $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, where $P(A)$ is the observed, and $P(E)$ is the expected probability. According to [6], $\kappa > 0.8$ indicates good reliability, and $0.67 < \kappa < 0.8$ is marginally reliable.

[6] Based on the questions labeled with a single type on which all three annotators agreed (434 questions in total).

ture work. The *Definition*, *Procedural*, and *Comparison* types constituted another significant group of questions (45.9%). Surprisingly, the *Causal* why-questions proved to be quite infrequent. To assess type distinguishability, we study agreement on individual question types following the procedure proposed by [9]. The lowest κ value is obtained for the *General Information Need* type. *General Information Need* type questions are underspecified since most of them are formulated as search queries using just a set of keywords, e.g. *"Mobile database management - design?"*

Question Type	**Frequency**	**Distinguishability (κ_{PO})**
Concept Completion	46.3%	.745
Definition	20.3%	.856
Procedural	17.1%	.803
Comparison	8.5%	.911
Causal	3.2%	.638
Disjunctive	1.8%	.702
Verification	1.4%	.756
Quantification	0.9%	.747
General Information Need	0.5%	.154

Table 2. Distribution and distinguishability of question types.

	PO	**PO$_{sure}$**	**CO**	**CO$_{sure}$**
A1-A3	.852	**.875**	.617	.780
A2-A3	.837	**.874**	.617	.789
A1-A2	.775	**.800**	.683	.738
A1-A1	**.947**	.900	.679	.867
A2-A2	.878	**.900**	.878	**.900**
A3-A3	.949	**1.0**	.772	**1.0**

Table 3. Inter-/intra-annotator κ on the question type.

Agreement on the question type annotation over all categories can be found in Table 3. The kappa value is reported for the four setups defined in Section 3.1. The upper part of the table presents inter-annotator agreement for 755 questions, the lower part corresponds to intra-annotator agreement for the 50 repeated questions. The examination of inter-annotator agreement shows that all methods of assessing kappa, apart from *CO*, yield reliable or marginally reliable agreement while the best results are obtained with PO_{sure}. The intra-annotator agreement shows that the annotation is stable, i.e. annotation results do not considerably vary over time.

Only 3.8% of all questions have been marked as opinion questions by all three annotators. We obtain low inter-annotator agreement[7] for this task which is caused by two major reasons: (i) correct decisions occasionally necessitate deep domain knowledge;

[7] Values for the pairwise agreement in opinion attribute annotation: κ_{A1-A3}=0.493, κ_{A2-A3}=0.396, κ_{A1-A2}=0.267

(ii) implicit requests for opinions can be too subtle to recognize. The opinion questions identified by all annotators are all explicit requests for opinions such as: *is it desirable to use technology to support teaching and learning in campuse-based courses?* We believe that a deeper study of opinion questions is needed in order to gain a better understanding of their properties.

The analysis of the question clarity attributes shows that about 1/5 of the questions are lexically, syntactically or semantically ill-formed[8]. In order to better understand the influence of question clarity on the manual question type classification, we measured inter-annotator agreement after removal of questions labelled with at least one question clarity attribute. The best inter-annotator agreement was obtained when ambiguous and syntactically ill-formed questions were removed. The surface-level ill-formedness caused by misspellings or Internet slang proved to be less detrimental to the question type annotation than ambiguity on the semantic level.

4. Conclusions

In this paper, we presented a question type classification scheme developed to gain a better understanding of the kinds of questions people ask on social Q&A sites. We used this scheme to annotate a sample of user questions and obtained good to very good inter-annotator agreement on this task. The annotation of opinion questions proved to be more difficult and hence necessitates further investigation. Around 1/5 of the questions were lexically, syntactically or semantically ill-formed. This observation has practical consequences for automatic QA systems aiming to deal with real and complex user questions: first, they have to integrate pre-processing components to handle surface level (lexical and syntactic) errors; second, they have to help users formulate better questions in order to get better answers.

Acknowledgements

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under the grant No. GU 798/3-1, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

[1] H. T. Dang, J. Lin, and D. Kelly, “Overview of the TREC 2006 Question Answering Track,” in *Proceedings of TREC 2006*, 2006.

[2] M. Tatham, “U.S. Visits to Question and Answer Websites Increased 118 Percent Year-over-Year.” [Online; visited March 26, 2008], March 19 2008. `http://www.hitwise.com/press-center/hitwiseHS2004/question-and-answer-websites.php`.

[3] A. Graesser, C. McMahen, and B. Johnson, “Question asking and answering,” in *Handbook of psycholinguistics* (M. Gernsbacher, ed.), ch. 15, pp. 517–538, San Diego: Academic Press, 1994.

[4] V. Stoyanov, C. Cardie, and J. Wiebe, “Multi-Perspective Question Answering Using the OpQA Corpus,” in *Proceedings of HTL-EMNLP 2005*, pp. 923–930, 2005.

[8] 16.8% are ambiguous, 20.1% are syntactically ill-formed, 8.7% contain Internet slang, 18.3% are misspelled.

[5] C. Müller and M. Strube, “Multi-level annotation of linguistic data with MMAX2,” in *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pp. 197–214, 2006.

[6] K. Krippendorff, *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc., 1980.

[7] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[8] A. Rosenberg and E. Binkowski, “Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points,” in *Proceedings of HLT-NAACL 2004*, pp. 77–80, 2004.

[9] S. Teufel, A. Siddharthan, and D. Tidhar, “An annotation scheme for citation function,” in *Proceedings of SIGDIAL-06*, (Sydney, Australia), 2006.