IT'S ANSWER TIME
Taking the Next Step in Question-Answering

Ana Cristina Mendes
L^{2}F/INESC-ID, IST Technical University of Lisbon, Lisbon, Portugal
ana.mendes@l2f.inesc-id.pt

Luísa Coheur
L^{2}F/INESC-ID, IST Technical University of Lisbon, Lisbon, Portugal
luisa.coheur@l2f.inesc-id.pt

Keywords: Answering, cooperative answer, correct, useful, non-misleading, answer generation.

Abstract: After all the work done in tasks like question classification, query expansion or information extraction in QA, we consider that some efforts should now be put specially on giving the answer to the user. In this paper we adopt the concept of cooperative answer – that is, a correct, useful and non-misleading answer – since it is our opinion that finding and presenting the cooperative answer to the user is one of the next challenges in QA. With that goal in mind, we focus on three main aspects that should deserve the attention of the QA community: the ability of systems to relate the candidate answers for a question; their ability to decide which candidates are possible final answers, given the question, but also the user who posed it; and, finally, the ability of generating the final answer in a cooperative way.

1 INTRODUCTION

Actual research on Question-Answering (QA) put much effort in tasks such as question classification (Li and Roth, 2002; Huang et al., 2008), query expansion (Bilotti et al., 2004; Derczynski et al., 2008) or information extraction from knowledge sources (Hovy et al., 2001; Frank et al., 2007). In what concerns answering, the used strategy involves the selection of the final answer from a list of extracted candidates from the involved texts. Usually, the final answer is the one with the higher score, which is typically calculated as a function of the candidate's frequency. As answering is a rather important phase in QA, we consider that this task needs urgent attention. In this paper we focus on three main research aspects that should be enhanced by QA community:

- **Relating answers**: QA systems usually base the process of recovering answers on redundancy (of the Web, or of document collections, even if in a smaller scale), that allows to directly extract answers based on the assumption that every information item is likely to have been stated multiple times, in multiple ways, in multiple documents (Lin, 2007). However, this redundancy-based strategy will only work if, in a first pace, a way of relating answers is devised. Some systems perform this step, as it is the case of the work by (Moriceau, 2005): prior to the selection phase, extracted answers are standardized. Nevertheless, a major challenge in QA research is still to be able to identify and relate groups of answers.
- **Targeting cooperative instead of correct answers**: current QA systems search after the answer that *correctly* solves the input question. Indeed, this characteristic marks a great difference between QA and Information Retrieval (IR) tasks: in the former the user does not need to filter and choose the answer from a set of possible answers, since the system has the ability, and responsibility, for doing that work for him. However, providing a *correct* answer is not enough, like (Gaasterland et al., 1992) concluded in their work on cooperative answering. Time has passed, research has progressed and systems have evolved, but the premises remain unchanged: better than a correct answer, is a *cooperative* answer, that is, a *correct, non- misleading and useful answer*. Nevertheless, the way how the previous three properties interconnect is far from trivial. A correct answer might not be useful, like an useful answer might not be

correct. Consider the question: *"How many kilobytes are there in a gigabyte?"*. For an IT student the useful answer would be the correct and exact one: *"1,048,576 KB"*; however, for someone who wants to have an idea of the magnitude of that amount, an useful answer could be *"1,048,000 KB"*, which is not correct. Here, the introduction of the adverb *around* could transform it in a correct and useful answer. Notice, too, that a correct answer can be misleading. Regard the question: *"How many Chechens did Stalin deport?"* and two possible answers *"more than 100,000"* and *"more than 100"*. Despite the correctness of both, the latter gives the user a wrong idea of greatness. If QA aims at being considered a valid and appealing alternative to IR (which has been adapting the retrieved results to the user), systems are required to follow this path. Even if finding correct answers is, by itself, a very hard task, which has certainly not yet seen its end, it does not suffice. Research must move forward in the direction of providing the right answer to the right user. YourQA took the first steps in this direction (Quarteroni and Manandhar, 2007), however the user has to first create his own profile.

- **Generating answers**: another aspect has to do with the formulation of the answer. The task is to create a valid and coherent answer, stated in natural language, if necessary. Directly related with this topic, one can not disregard the amount of work done in Natural Language Generation (NLG) (Reiter and Dale, 2000). Nevertheless, and despite the deep interest on many other sub-tasks of QA systems, answer generation is still a topic to be grubbed: the usual strategy is to return directly what was extracted. (Moriceau, 2005) is an exception, with an approach that computes and lexicalizes the answer to be given to the user.

This paper is organized as follows: in Section 2 we present a typology of relations and survey work done in this direction. In Section 3 we briefly survey some of the community efforts towards cooperative answers and in Section 4 we focus on the problematic of answer generation. The paper concludes in Section 5.

2 RELATING ANSWERS

Relating candidate answers is a challenge to be solved by QA systems. Here we present a typology of relations that connect answers, which we derived from the analysis of corpora with questions and answers (Magnini et al., 2003; Magnini et al., 2005), and based on the terminology described in (Moriceau, 2006), firstly proposed by (Webber et al., 2002). Moreover, we discuss how groups of answers can be identified.

2.1 A typology of relations

Relations between candidate answers can be of:

EQUIVALENCE: if answers are consistent and entail mutually, namely:

1) answers with notational variations. For instance, *"Oct. 14, 1947"* and *"14th October, 1947"* are equivalent answers for *"When did the test pilot Chuck Yeager break the sound barrier?"*;

2) answers that rephrase others, from synonyms to paraphrases. The question *"How did Jack Unterweger die?"* can be answered with *"committed suicide"* or *"killed himself"*;

INCLUSION: if answers are consistent and differ in specificity, one entailing the other, through:

1) hypernymy: *"What animal is Kermit?"* can be answered with *"frog"* or *"amphibian"*; or

2) meronymy: for instance, *"Where did Ayrton Senna have the accident that caused his death?"*, in which *"Imola"*, *"Italy"*, *"Europe"* and *"earth"* are possible answers; or

3) membership: for instance, *"Edvard Munch"* and *"a Norwegian Symbolist painter"* are both possible answers to *"Who painted the "Scream"?"*

ALTERNATIVE: if answers are not entailing:

1) representing distinct and complementary visions of the same entity. For example, *"the largest wind energy park of Hessia"*, and *"six wind turbines with a power of 1350 kilowatts"* are two complementary answers for *"What is set up in Seibertenrod in Vogelsberg?"*; or,

2) representing different entities, and can be used together by means of a conjunction. *"Aribert Heim"*, *"Josef Mengele"* and *"Jack Kevorkian"* are answers for *"What is the real name of Dr. Death?"*, used separately or in a conjunction.

CONTRADICTORY: if answers are inconsistent and their conjunction is invalid. The question *"How is the Pope?"* can be answered with *"ill"* or *"healthy"*, but not with both.

2.2 Identifying groups of answers

String distance metrics – like the Levenshtein distance, the cosine or Jaccard similarities – can be used to relate equivalent answers: *"John Kennedy"* and *"John F. Kennedy"* are equivalent, as they (most of the times) refer to the same person. However, this approach has some limitations, since it can not be always directly applied. For example, in *"George Bush"* and *"George W. Bush"*, it depends on the context whether they refer to the same person or to two different persons.

Normalization is another strategy to encounter equivalence relations, aiming to find canonical unambiguous referent for entities. (Moriceau, 2005) deals with diversity in candidate answers, and presents an approach to deliver single coherent answers to `date` questions. The same author (Moriceau, 2006) described a method to deal with variation in numerical answers, in which the use of frames containing all information related with numerical values allow comparisons between answers. Other experiments (Khalid et al., 2008) in named entity normalization have shown that it helps text retrieval for QA. Normalization is typically done prior to the answer extraction phase, and does not aim at connecting answers through equivalence after they were extracted. Like in paraphrase detection, it is used to permit the extraction of diverse candidate answers referring to the same entity (Takahashi et al., 2003; France et al., 2003).

Inclusion and equivalence relations can be built by using the lexical relations present in Wordnet[1]. This approach was used by (Dalmas and Webber, 2007), who propose a technique to organize answer candidates on the geographical domain into clusters. Answers' models are created from questions and their candidates, and represented as direct graphs expressing the fusion of information contained in the set of extractions. The final answer is retrieved based on the computation of properties that compare the graphs nodes.

The detection of alternative and contradictory relations require different procedures. The former deals with discovering if two answers point to the same entity; but, in contrary to equivalence, inference based on lexical relations is not enough. In the latter, several notions should be taken into account, namely: if the answers are antonyms (Mohammad et al., 2008); the quality and trustworthiness of the document in which the answer was found (Oh et al., 2009), the time period answers refer to (specially important if they are searched in the Web).

3 COOPERATIVE ANSWER *VERSUS* CORRECT ANSWER

In order to choose the cooperative answer to a given question, the procedure should be taken further than uniquely discovering which candidates are correct. This decision depends, at great extent, on the one who will get the answer: the user. Here we discuss the properties that make an answer cooperative: being correct, non-misleading and useful.

3.1 Correct answers

The correctness of an answer relates, firstly, with whether it is associated with the entity that the question is seeking after. For instance, the question *"What animal is Kermit?"* asks for the character in The Muppets Show, and not the computer protocol. The focus is, thus, to identify which answers hold contradictory relations and, from those, to decide which to choose. Most QA systems consider this as a main and final goal: to select the correct answer among all candidates.

3.2 Non-misleading answers

A non-misleading answer avoids the user to create a wrong interpretation about the topic under consideration.

Here we focus on three problems that can mislead the user when dealing with an open-domain QA system over large collections of text: 1) answer ambiguity, 2) answer granularity, and 3) answer absence.

- Ambiguity, as it is usually considered in QA, arises either from corpora sources or user questions. Systems that do try to cope with ambiguity in the user's question, usually push its resolution to the user side, through the use of *clarification dialogues*. Ambiguity in answers, however, is usually not addressed. Returning an ambiguous answer to a question is not cooperative, since it leaves room to multiple interpretations. For instance, answering *Bush* to the question *"Who was the President of the US during the Gulf War?"* is ambiguous since there were two presidents of the United States named Bush, and the answer does not clearly state each one responds to the question.
- The problem with answer granularity resides in the fact that, in many situations, the decision about which answer to choose for a question is fuzzier: also among human assessors there is no agreement about what is the answer to a given question. (Lin and Katz, 2006) mention that granularity is a critical point specially if the answer belongs to the types PERSON, LOCATION or DATE. Till a certain point, the granularity of the cooperative answer depends on the question, and on several other factors external to the questioner (like

[1] `http://wordnet.princeton.edu/`

its position in space (Shanon, 1979) and, by analogy, in time) and can thus be controlled with recourse to rules and guidelines. On the other hand, it is certainly not independent from him, his characteristics and what he expects to have as answer. As (Lin and Katz, 2006) point out, the granularity has much to do with real users: 'better understanding of real-world user needs will lead to more effective question answering systems in the future'.

- Answer absence has to do with responding to the user when no answer was found. Indeed, and albeit a wrong answer contradicts the goal of QA, retrieving no answer is not better: it does not bring any valuable information to the user, and can lead him/her to misinterpretations. Besides originating on the side of the system – which could not find an answer within the available corpora – questions with no answer can arise from the user side, namely from false presuppositions. Consider the question *"Who is the King of France?"*. Knowing that France is a republic, answering *"no one"* (or, even worse, *"NIL"*) can drive the user to think the system was unable to find the answer. In this case, an explanation is due. (Benamara, 2004) reports WEBCOOP, a restricted domain logic-based system that integrates knowledge representation and advanced reasoning to detect false presuppositions and misunderstandings in questions, in order to deliver non-misleading answers. To our knowledge, there is no open domain QA system that deals with this problematic.

3.3 Useful answers

Choosing the useful answer is a task deeply intertwined with the characteristics of the questioner. There is, thus, the need for acknowledging the user and his/her goals. We consider that it can be achieved through the recognition of the clues the user provides when interacting with the system, in three different levels: question, context and history of interactions.

Question clues: The first level in finding clues for deciding which answer is useful is to analyze the question and how it was posed to the system. Several clues in the question can be checked, namely (and surely not limited to):

1) the vocabulary, that differs depending on the user. For instance, on his/her age, academic background and occupation;

2) the specificity and world knowledge. If the user asks: *"Who received the Prince of Asturias Award for Technical and Scientific Research for his studies on the discovery of the first synthetic vaccination against malaria?"*, probably he has deep knowledge on the topic.

DuARTE Digital (Mendes et al., 2009) answers questions about a piece of jewellery, and dynamically tries to assess its interlocutor characteristics, based on the used vocabulary. With a list of words that naive or expert users might employ, it interprets, at every question, the user's expertise. Knowing this, it chooses the answer (previously marked with the corresponding difficulty level) from a knowledge base.

Clues in the current context: The next level has to do with the current context. Consider a chain of questions about the 2^{nd} World War. If the question *"What is the real name of Dr. Death?"* appears, probably the useful answer will be *"Aribert Heim"* or *"Josef Mengele"* (or both) and not any other from the set of possible answers[2].

Again, DuARTE Digital (Mendes et al., 2009) is a system that tries to acknowledge the user's goals at a contextual level. It measures the proximity of the user's words in a sequence of questions to different sub-topics, in order to understand the orientation of the interaction. By doing so, it distinguishes from focused to stray interactions, and chooses the answer according to its detail and informative level.

Clues in the history of interactions: The final level relates with the history of interactions between user and system.

Although this is not a new issue in IR, specially in Web search engines, where systems try to adapt the presentation of results according to the user (Liu et al., 2004; Teevan et al., 2005), the first steps in QA only recently have been taken. (Quarteroni and Manandhar, 2009) were pioneers in this topic, as they included on the system YourQA a component dedicated to user modelling. The purpose is to filter the documents where answers are searched and to rerank the candidates based on the degree of match with the user's profile. Users must create their own profile when first interacting with the system (it does not dynamically discovers its interlocutor characteristics), and their browsing history is taken in consideration in future interactions. Any question submitted to the system is answered by taking the user's profile into account.

[2] Notice the multiple candidate answers for this question: http://en.wikipedia.org/wiki/Dr._Death

4 ANSWER GENERATION

The simplest approach that can be envisioned is to return the most frequent candidate. Returning the complete set of correct answers can also be an option, like for *"What is the real name of Dr. Death?"*. On the other hand, some answers can be incorporated on a single one that subsumes the set. For instance, *animal* can be chosen to answer the question *"What is Kermit?"*, instead of *frog* or *amphibian*. Nevertheless, here the point is common: these are answers based on pure extraction, and not their generation.

The fact is, although much work as been devoted to NLG, it seems that QA still could not benefit from the results achieved in this well-established field.

Answer generation is preferable to answer extraction for the purpose of answering: firstly it *humanizes* the system; second, it permits the usage of adapted vocabulary; finally, it allows the introduction of information that the user did not explicitly request, but might be interested in.

There are a few examples of works that try to *build* answers, instead of merely extract and retrieve. Again, (Moriceau, 2005)'s work in data integration is a good contribution. The system generates natural language answers by making use of generation schemas, and lexicalizes its degree of confidence in the answer with the use of adverbs and their intensities (*e.g.*, possibly, most possibly, probably, most probably). Another example is WEBCOOP (Benamara and Saint-Dizier, 2003), which relies on templates to report user misconceptions and display solutions to help to the user, in natural language.

5 CONCLUSION

In this paper we introduced our position about what we believe to be a desirable focus of research in QA: the process of retrieving the answer to the user. We presented three main aspects that should lead the research in QA, namely: relating the candidate answers for a given question and for that we proposed a typology of four relations; choosing the answer among the candidates, rather than only the correct one; and, generating the final answer, by using the work already done in NLG, instead of retrieving the extracted information chunks.

There are a few systems that partly cope with some of the presented problems. We made a brief survey on these systems. Nevertheless, to our knowledge there is no open-domain QA that deals with each of these as a whole, for the purpose of answering. We consider that the goal is now to put them together, and focus on QA systems and approaches that take the problem of cooperative answering into consideration. Systems should evolve in this direction, to become more competitive and appealing to real end-users.

ACKNOWLEDGEMENTS

Ana Cristina Mendes is supported by a PhD fellowship from Fundação para a Ciência e a Tecnologia (SFRH/BD/43487/2008).

REFERENCES

Benamara, F. (2004). Cooperative question answering in restricted domains: the webcoop experiment. In *Proc. ACL Workshop on Question Answering in Restricted Domains*, pages 21–26.

Benamara, F. and Saint-Dizier, P. (2003). Dynamic generation of cooperative natural language responses. In *EWNLG03-ACL , Budapest,* , pages 56–67. ACL.

Bilotti, M., Katz, B., and Lin, J. (2004). What works better for question answering: Stemming or morphological query expansion. In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*.

Dalmas, T. and Webber, B. L. (2007). Answer comparison in automated question answering. *Journal of Applied Logic*, 5(1):104–120.

Derczynski, L., Wang, J., Gaizauskas, R., and Greenwood, M. (2008). A data driven approach to query expansion in question answering. In *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pages 34–41. Association for Computational Linguistics.

France, F. D., Yvon, F., and Collin, O. (2003). Learning paraphrases to improve a question-answering system. In *Proc. 10th Conference of EACL Workshop Natural Language Processing for Question-Answering*.

Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jrg, B., and Schfer, U. (2007). Question answering from structured knowledge sources. *Journal of Applied Logic*, 5(1):20 – 48. Questions and Answers: Theoretical and Applied Perspectives.

Gaasterland, T., Godfrey, P., and Minker, J. (1992). An overview of cooperative answering. *Journal of Intelligent Information Systems*, 1(2):123–157.

Hovy, E., Hermjakob, U., and yew Lin, C. (2001). The use of external knowledge in factoid qa. In *In Proceedings of the Tenth Text REtrieval Conference (TREC*, pages 644–652.

Huang, Z., Thint, M., and Qin, Z. (2008). Question classification using head words and their hypernyms. In *EMNLP*, pages 927–936.

Khalid, M., Jijkoun, V., and de Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. In *30th European Conference on Information Retrieval (ECIR 2008)*, pages 705–710. Springer, Springer.

Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Lin, J. (2007). An exploration of the principles underlying redundancy-based factoid question answering. *ACM Trans. Inf. Syst.*, 25(2):6.

Lin, J. and Katz, B. (2006). Building a reusable test collection for question answering. *J. Am. Soc. Inf. Sci. Technol.*, 57(7):851–861.

Liu, F., Yu, C., and Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Penas, A., Peinado, V., Verdejo, F., de Rijke, M., and Vallin, R. (2003). The multiple language question answering track at clef 2003. In *CLEF 2003. CLEF 2003 Workshop*. Springer-Verlag.

Magnini, B., Vallin, R., Ayache, C., Erbach, G., Peñas, A., Rijke, M. D., Rocha, P., Simov, K., and Sutcliffe, R. (2005). Overview of the clef 2004 multilingual question answering track. In *Results of the CLEF 2004 Cross-Language System Evaluation Compaign*, pages 371–391. Springer-Verlag, Berlin Hidelberg.

Mendes, A. C., Prada, R., and Coheur, L. (2009). Adapting a virtual agent to users vocabulary and needs. In *Proc. 9th International Conference IVA 2009*, LNAI 5773. Springer-Verlag.

Mohammad, S., Dorr, B., and Hirst, G. (2008). Computing word-pair antonymy. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 982–991, Morristown, NJ, USA. Association for Computational Linguistics.

Moriceau, V. (2005). Answer generation with temporal data integration. In Wilcock et al., G., editor, *EWNLG, Aberdeen*, pages 197–202. University of Aberdeen.

Moriceau, V. (2006). Numerical data integration for cooperative question-answering. In *Knowledge and Reasoning for Language Processing (KRAQ), Trento, Italy*, pages 43–50. Association for Computational Linguistics (ACL).

Oh, H.-J., Lee, C.-H., Yoon, Y.-C., and Jang, M.-G. (2009). Question answering based on answer trustworthiness. In *Information Retrieval Technology, 5th Asia Information Retrieval Symposium, AIRS 2009, Sapporo, Japan, October 21-23, 2009. Proceedings*, pages 310–317.

Quarteroni, S. and Manandhar, S. (2007). User modelling for personalized question answering. In *AI*IA '07: Proc. 10th Congress of the Italian Association for Artificial Intelligence on AI*IA 2007*, pages 386–397, Berlin, Heidelberg. Springer-Verlag.

Quarteroni, S. and Manandhar, S. (2009). Designing an interactive open-domain question answering system. *Natural Language Engineering*, 15(1):73–95.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Natural Language Processing. Cambridge University Press.

Shanon, B. (1979). Where questions. In *Proceedings of the 17th annual meeting on Association for Computational Linguistics*, pages 73–75, Morristown, NJ, USA. Association for Computational Linguistics.

Takahashi, T., Nawata, K., Inui, K., and Matsumoto, Y. (2003). Effects of structural matching and paraphrasing in question answering (special issue on text processing for information access). *IEICE transactions on information and systems*, 86(9):1677–1685.

Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *SIGIR '05: Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456. ACM.

Webber, B., Gardent, C., and Bos, J. (2002). Position statement: Inference in question answering. In *Proc. 3rd international conference on language resources and evaluation (LREC)*, pages 19–25.