Language Resources and Evaluation

Analyzing the Capabilities of Crowdsourcing Services for Text Summarization

--Manuscript Draft--

Manuscript Number:	LREV1124R1
Full Title:	Analyzing the Capabilities of Crowdsourcing Services for Text Summarization
Article Type:	Full length article, original research
Keywords:	Information retrieval; Text Summarization; Crowdsourcing services; Crowdflower; Mechanical Turk
Corresponding Author:	Elena Lloret, Ph.D. University of Alicante Alicante, SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Alicante
Corresponding Author's Secondary Institution:	
First Author:	Elena Lloret, Ph.D.
First Author Secondary Information:	
Order of Authors:	Elena Lloret, Ph.D.
	Laura Plaza
	Ahmet Aker
Order of Authors Secondary Information:	
Manuscript Region of Origin:	
Abstract:	This paper presents a detailed analysis of the use of crowdsourcing services for the Text Summarization task in the context of the tourist domain. In particular, our aim is to retrieve relevant information about a place or an object pictured in an image in order to provide a short summary which will be of great help for a tourist. For tackling this task, we proposed a broad set of experiments using crowdsourcing services that could be useful as a reference for others who want to rely also on crowdsourcing. From the analysis carried out through our experimental setup and the results obtained, we can conclude that although crowdsourcing services were not good to simply gather gold-standard summaries (i.e., from the results obtained for experiments 1, 2 and 4), the encouraging results obtained in the third and sixth experiments motivate us to strongly believe that they can be successfully employed for finding some patterns of behaviour humans have when generating summaries, and for validating and checking other tasks. Furthermore, this analysis serves as a guideline for the types of experiments that might or might not work when using crowdsourcing in the context of text summarization.
Response to Reviewers:	Dear Editor and Reviewers, This is the report of how we have taken into account the comments by the two reviewers regarding our original submission (Submission LREV1124) to the Language Resources and Evaluation journal. Comments extracted from the reviews are presented as headings, with the corresponding explanations following them. # Reviewer 1 1. Overall the findings are not too surprising. It's good to know the results, although there is not much novelty in the study. The main aim in this study was to provide a in-depth analysis through a broad

experimental framework of how crowdsourcing services work within the multi-document text summarization task. So far, most of the research works involving crowdsourcing show very good results, and state that this type of services are useful. However, to the best of our knowledge there is no previous work that analyzes the reliability of crowdsourcing for multi-document summarization, and more specifically, in the context of retrieving relevant information regarding a place or an object represented by an image, pertaining to the tourist domain.

2. In abstract, I feel some conclusions are too general. For example, "we observed that this type of services may be helpful for specific natural language processing tasks when they are easy and enjoyable (...), but not when the tasks involve much reading and understanding.".

The abstract has been changed in order to avoid too general statements.

3. Also the statement in abstract -- the easier the task, the higher the chances to obtain malicious responses -- seems not well supported by the study.

As it was stated in the previous comment, the abstract has been changed in order to avoid too general statements.

4. Using crowdsourcing for text summarization, there are two places that need human annotation -- writing human summaries, and evaluating system generated summaries. I think the authors should make it clear what their study is about early on in the paper.

This has been clarified in the paper (see Section Introduction):
Therefore, the main objective of this paper is to analyze to what extent crowdsourcing services are useful for text summarization, in particular, for the task of writing human summaries that may be use as gold-standard or model summaries for evaluation.

5. The ROUGE scores are not very sensitive. >From the results in Table 2 (and other similar tables), it is hard to say what the quality of the AMT summaries is. The lower scores can mean lower quality, but not necessarily garbage, with malicious errors generated by the annotators mentioned in the paper.

We compute ROUGE results in order to have an idea of how the summaries perform with respect to the model ones ones. We assume that, since human annotators are selecting relevant sentences from documents, these sentences will be similar of that of model summaries. This has been made clear in the paper.

6. The number of annotators is very different for different experiments. I'm wondering why that is the case.

The number of annotators differs across experiments for several reasons. In some experiments (e.g., experiments 2 and 4 using trap sentences), in the light of the poor quality of the partial results, we increased the number of annotators. Other experiments (e.g., experiment 3) were more simple and showed better quality results, and so less annotators were needed. In experiment 6, given that the number of summaries to evaluate was not big, we estimated that 4 annotators would be enough for our purposes.

7. I feel sometimes the conclusions are too strong, just based on the analysis of one set of annotation results.

Sentences like these have been changed.

8. In experiment 2, there is some difference between its results and experiment 1. Does adding trap sentences in experiment 2 explain the difference? I don't think so. The annotators are making a lot of random errors (not paying attention to the task). The authors found that annotators didn't select the first sentences as often as in experiment 1. I don't think that is a real pattern, rather it seems it is just randomness among annotators.

In this experiment, despite including “trap sentences” to control the performance of the annotators, the results were not satisfactory either. We have rewritten the suggested sentence.

9. For this experiment, the authors observe better ROUGE score for annotation 2, but that doesn't really mean anything -- that's just the best annotator. This can't be used to demonstrate that adding trap sentences helps annotation quality (for that the overall/average results need to better). Even for experiment 1, if the best annotation is selected, its result might be quite good too.

As we state in the conclusions for Experiment 2, the results obtain and the summaries generated were not good, as it occurred with Experiment 1. The addition of trap sentences, may not have positively influence on the annotators, who still were performing the task randomly. That is why we decide to narrow the task in the remaining experiments. This is been clarified and better explained in the paper.

10. Experiment 3 is not really a summarization task. This is the only setup that result in good annotation quality.

Although this experiment does not directly involves a summarization task, we believe it is important, since it is a valuable indicator that can be very useful in the selection of relevant sentences for the summary, as it is done in further experiments. This has been clarified in the paper.

11. For Experiment 4, the authors choose not to use ROUGE since the quality of annotation is bad. But ROUGE is used in previous experiments which also have poor annotation quality. The experiment setup should be better justified.

The reason why we do not use ROUGE in this experiment has been better justified. On the one hand, the length of the resulting summaries was not comparable to the model ones, and consequently, comparing them using ROUGE would have produced distort results with incorrect interpretations. On the other hand, we did not know a priori, whether the documents contain the answer to the proposed questions, so from out point of view, this type of analysis was more interesting.

12. In experiment 5, I'm wondering whether the reference/model summaries is appropriate. The summaries are generated based on the 5 questions thought by the annotators, but the model summaries are more generic.

Yes, that is an important point. In our previous work (Do humans have conceptual models about Geographic Objects? A user study) we have shown that humans have conceptual models about what types of information (e.g. location information about an object) to seek about locations. Our model summaries include the types of information determined as relevant for a location. The questions collected in experiment 3 reveal also that the MTurk workers seek similar information types as included in the model summaries. In experiment 5 we used the experiment 3 as basis and ask the workers to think about a set of questions (as done in experiment 3) and find the answers within the documents. Note that the documents contained answers for questions collected through the experiment 3. If the workers in experiment 5 followed the task properly we are sure that the overlap between their answers and the model summaries were reasonable high and thus we think that our model summaries are appropriate for judging the automatic summaries.

Aker A, Plaza L, Lloret E. Do humans have conceptual models about Geographic Objects? A User Study. Journal of the American Society for Information Science and Technology (JASIST). In press.

13. The authors said that using control mechanism may be helpful to obtain summaries, that is not supported by the results.

We have clarified in the paper that, even though the use of quality control mechanisms may be of help in some cases, in general they are not a guarantee that the MTurks are not cheating.

[Conclusion and Future Work] From experiments 5 and 6, we can also conclude that the quality control mechanisms introduced may be of help in some cases, although they are not a guarantee that the annotators will be committed to the task, as we noticed in experiment 5, where the results were not very good either.

[Experiment 5] The results of this experiment seem to indicate that the use of control mechanisms may be of help when obtaining automatic summaries from crowdsourcing services, since they allow to quickly detect malicious user behavior. However, this control policy does not seem to guarantee the quality of the annotations, and still the results are not satisfactory.

14. Experiment 6 is more about evaluating summaries. The data set (after removing some poor annotations) is rather small for meaningful conclusions.

We do not agree with the reviewer in this point. We think that the set of 33 summaries is big enough to get significant results. According to the study presented in (Lin, 2004) this number is large enough to ensure significant results in a single document evaluation task using ROUGE metrics.

Lin CY: Looking for a few good metrics: Automatic summarization evaluation - How many samples are enough? In Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization 2004.

15. In general, I think the paper makes stronger conclusions than what's supported by the experimental results. There is a lot of variation in experimental setups (annotators, task definition, postprocessing, etc), the authors may have generalized too much based on specific observations.

The strong claims and statements in the paper have been rewritten. The whole paper has been modified in order to better justify the structure and organization of the different experiments performed. Moreover, general sentences have been also changed into more specific ones, depending on each experiment and result.

# Reviewer 2

1. While the experiments are clearly presented and logically structured, there is some missunderstanding when presenting the ROUGE comparisons. Namely, the experimental description can be interpreted as if the authors compared the AMT summaries first against the manually built model and then against the summary corpus derived from Wikipedia (e.g., on p14: " Table 4 shows a comparison between these [AMT] summaries with respect to the model summaries and the Wikipedia baseline"). Yet, none of the tables 2, 4, 11, 14 have a row for the comparison of AMT summaries with Wikipedia (the Wikipedia corpus is only compared against the model). Can the authors please clarify this?

There was a mistake in the sentence on page 14. It has been corrected. For all expertiments using ROUGE as evaluation tool, the AMT summaries were compared only to the model ones. We did not perform a direct comparison between AMT summaries and Wikipedia summaries. However, we did compare the results of the Wikipedia summaries with respect to the model ones. The reason why doing this (model vs. model and wikipedia vs. model) was to have an idea of a upper bound and analyze how far our results were from them.

2. On top of p6, authors should also include a reference to the third author's work, just published at LREC [1], in which their findings about the influence of payment on quality contradict those of Mason and Watts.

[1] Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour and Udo Kruschwitz. assessing Crowdsourcing Quality through Objective Tasks. LREC'12.

Yes, we included the reference.

3. Related to the above, another intriguing aspect that is not discussed is how the payment value could have interfered with the results. For example, in experiment 5 the

author payed .15 per HIT (which is on average the triple of the amount payed in the other experiments) and as a result this batch of tasks was also completed in the shortest time span (only 2 hours). Yet the time spent by turkers on individual tasks is very low, most of the tasks being completed in an unreasonable time span of 0.5 seconds!! Is this yet another proof that too high payments attract too many cheaters? A discussion and a comparison between the payment, completion time, average task solving time and the overall result quality could be very interesting in the concluding section. Also, how does this compare against one of the author's findings in [1], where a positive correlation between payment and quality was found?

[1] Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour and Udo Kruschwitz. assessing Crowdsourcing Quality through Objective Tasks. LREC'12.
We included this into the conclusion:

Finally, the results seem to indicate that the observed low reliability may be due to motivational factors, and this aspect should be studied in future work. In previous work payment was studied as one of the motivational factor for controlling the results quality. Aker et al. (2012) showed that high payments lead to better results whereas in Mason and Watts (2010) and Feng et al., (2009) it is discussed that an increase in payment attracts more spammers and as consequence it leads to low quality results. In our summary generation experiments we varied the payments in small steps from low to high however, in overall we have not manage to obtain any useful results from the MTurk workers. As discussed above the only useful results were obtained through the experiment 3. This experiment 3 differed from the other in a way that it was far easy to complete and did not require any time consuming reading task. Thus, we think that the major factor in obtaining high quality results is the level of difficulty of a task. If a task is easy and fast to perform then we think that this will positively influence the results quality. On ther other hand if the task requires a lot of time to complete and is difficult to perform this causes that the workers loose motivation on the task which then reflects the results negatively.

4. Typos:

All typos and spelling errors have been corrected. Moreover, the paper has been proofread in order to avoid this type of errors.

p2: have been analyze => has been analyzed → Done
p4: broadly categorize => broadly categorized → Done
p4: $60 USD => $60 (also, pleas be consisten through the document, e.g., in some places you have "0.05 US dollars" → Done. All the quantities have been unified into US dollars
p17: not time enough => not enough time → Done
p18: confirms us that => confirms that → Done
p20: we broad => we broaden → Done
p22, caption Fig.5: th euse of the "image annotation" term does not reflect the experiment itself → This has been changed.
p24: second sentence of Section 9.2 does not make sense → This sentence has been removed.
p25: do not obtained => did not obtain → Done
p27: Crowdflower through AMT => AMT through Crowdflower → Done
p27: even being aware => even if being aware → Done
p27: higher in complexity => more complex → Done
p28: guideles as far as types of experiments that should and should not work => guidelines such as types of experiments that might or might not work → Done

Best regards,
Elena Lloret
University of Alicante

Report on corrections for LREV journal article 01/08/2012

Elena Lloret
Laura Plaza
Ahmet Aker

This is the report of how we have taken into account the comments by the two reviewers regarding our original submission (Submission LREV1124) to the Language Resources and Evaluation journal. Comments extracted from the reviews are presented as headings, with the corresponding explanations following them.

Reviewer 1

1. **Overall the findings are not too surprising. It's good to know the results, although there is not much novelty in the study.**

The main aim in this study was to provide a in-depth analysis through a broad experimental framework of how crowdsourcing services work within the multi-document text summarization task. So far, most of the research works involving crowdsourcing show very good results, and state that this type of services are useful. However, to the best of our knowledge there is no previous work that analyzes the reliability of crowdsourcing for multi-document summarization, and more specifically, in the context of retrieving relevant information regarding a place or an object represented by an image, pertaining to the tourist domain.

2. **In abstract, I feel some conclusions are too general. For example, "we observed that this type of services may be helpful for specific natural language processing tasks when they are easy and enjoyable (...), but not when the tasks involve much reading and understanding.".**

The abstract has been changed in order to avoid too general statements.

3. **Also the statement in abstract -- the easier the task, the higher the chances to obtain malicious responses -- seems not well supported by the study.**

As it was stated in the previous comment, the abstract has been changed in order to avoid too general statements.

4. **Using crowdsourcing for text summarization, there are two places that need human annotation -- writing human summaries, and evaluating system generated summaries. I think the authors should make it clear what their study is about early on in the paper.**

This has been clarified in the paper (see Section Introduction):
Therefore, the main objective of this paper is to analyze to what extent crowdsourcing services are useful for text summarization, in particular, for the task of writing human summaries that may be use as gold-standard or model summaries for evaluation.

5. **The ROUGE scores are not very sensitive. >From the results in Table 2 (and other similar tables), it is hard to say what the quality of the AMT summaries is. The lower scores can mean lower quality, but not necessarily garbage, with malicious errors generated by the annotators mentioned in the paper.**

We compute ROUGE results in order to have an idea of how the summaries perform with respect to the model ones ones. We assume that, since human annotators are selecting relevant sentences from documents, these sentences will be similar of that of model summaries. This has been made clear in the paper.

6. **The number of annotators is very different for different experiments. I'm wondering why that is the case.**

The number of annotators differs across experiments for several reasons. In some experiments (e.g., experiments 2 and 4 using trap sentences), in the light of the poor quality of the partial results, we increased the number of annotators. Other experiments (e.g., experiment 3) were more simple and showed better quality results, and so less annotators were needed. In experiment 6, given that the number of summaries to evaluate was not big, we estimated that 4 annotators would be enough for our purposes.

7. **I feel sometimes the conclusions are too strong, just based on the analysis of one set of annotation results.**

Sentences like these have been changed.

8. **In experiment 2, there is some difference between its results and experiment 1. Does adding trap sentences in experiment 2 explain the difference? I don't think so. The annotators are making a lot of random errors (not paying attention to the task). The authors found that annotators didn't select the first sentences as often as in experiment 1. I don't think that is a real pattern, rather it seems it is just randomness among annotators.**

In this experiment, despite including “trap sentences” to control the performance of the annotators, the results were not satisfactory either. We have rewritten the suggested sentence.

9. **For this experiment, the authors observe better ROUGE score for annotation 2, but that doesn't really mean anything -- that's just the best annotator. This can't be used to demonstrate that adding trap sentences helps annotation quality (for that the overall/average results need to better). Even for experiment 1, if the best annotation is selected, its result might be quite good too.**

As we state in the conclusions for Experiment 2, the results obtain and the summaries generated were not good, as it occurred with Experiment 1. The addition of trap sentences, may not have positively influence on the annotators, who still were performing the task randomly. That is why we decide to narrow the task in the remaining experiments. This is been clarified and better explained in the paper.

10. **Experiment 3 is not really a summarization task. This is the only setup that result in good annotation quality.**

Although this experiment does not directly involves a summarization task, we believe it is important, since it is a valuable indicator that can be very useful in the selection of relevant sentences for the summary, as it is done in further experiments. This has been clarified in the paper.

11. **For Experiment 4, the authors choose not to use ROUGE since the quality of annotation is bad. But ROUGE is used in previous**

experiments which also have poor annotation quality. The experiment setup should be better justified.

The reason why we do not use ROUGE in this experiment has been better justified. On the one hand, the length of the resulting summaries was not comparable to the model ones, and consequently, comparing them using ROUGE would have produced distort results with incorrect interpretations. On the other hand, we did not know a priori, whether the documents contain the answer to the proposed questions, so from out point of view, this type of analysis was more interesting.

12. **In experiment 5, I'm wondering whether the reference/model summaries is appropriate. The summaries are generated based on the 5 questions thought by the annotators, but the model summaries are more generic.**

Yes, that is an important point. In our previous work (Do humans have conceptual models about Geographic Objects? A user study) we have shown that humans have conceptual models about what types of information (e.g. location information about an object) to seek about locations. Our model summaries include the types of information determined as relevant for a location. The questions collected in experiment 3 reveal also that the MTurk workers seek similar information types as included in the model summaries. In experiment 5 we used the experiment 3 as basis and ask the workers to think about a set of questions (as done in experiment 3) and find the answers within the documents. Note that the documents contained answers for questions collected through the experiment 3. If the workers in experiment 5 followed the task properly we are sure that the overlap between their answers and the model summaries were reasonable high and thus we think that our model summaries are appropriate for judging the automatic summaries.

Aker A, Plaza L, Lloret E. Do humans have conceptual models about Geographic Objects? A User Study. Journal of the American Society for Information Science and Technology (JASIST). In press.

13. **The authors said that using control mechanism may be helpful to obtain summaries, that is not supported by the results.**

We have clarified in the paper that, even though the use of quality control mechanisms may be of help in some cases, in general they are not a guarantee that the MTurks are not cheating.

[Conclusion and Future Work] *From experiments 5 and 6, we can also conclude that the quality control mechanisms introduced may be of help in some cases, although they are not a guarantee that the annotators will be committed to the task, as we noticed in experiment 5, where the results were not very good either.*

[Experiment 5] *The results of this experiment seem to indicate that the use of control mechanisms may be of help when obtaining automatic summaries from crowdsourcing services, since they allow to quickly detect malicious user behavior. However, this control policy does not seem to guarantee the quality of the annotations, and still the results are not satisfactory.*

14. **Experiment 6 is more about evaluating summaries. The data set (after removing some poor annotations) is rather small for meaningful conclusions.**

We do not agree with the reviewer in this point. We think that the set of 33 summaries is big enough to get significant results. According to the study presented in (Lin, 2004) this number is large enough to ensure significant results in a single document evaluation task using ROUGE metrics.

Lin CY: Looking for a few good metrics: Automatic summarization evaluation - How many samples are enough? In Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization 2004.

15. **In general, I think the paper makes stronger conclusions than what's supported by the experimental results. There is a lot of variation in experimental setups (annotators, task definition, postprocessing, etc), the authors may have generalized too much based on specific observations.**

The strong claims and statements in the paper have been rewritten. The whole paper has been modified in order to better justify the structure and organization of the different experiments performed. Moreover, general sentences have been also changed into more specific ones, depending on each experiment and result.

Reviewer 2

1. **While the experiments are clearly presented and logically structured, there is some missunderstanding when presenting the ROUGE comparisons. Namely, the experimental description can be interpreted as if the authors compared the AMT summaries first against the manually built model and then against the summary corpus derived from Wikipedia (e.g., on p14: " Table 4 shows a comparison between these [AMT] summaries with respect to the model summaries and the Wikipedia baseline"). Yet, none of the tables 2, 4, 11, 14 have a row for the comparison of AMT summaries with Wikipedia (the Wikipedia corpus is only compared against the model). Can the authors please clarify this?**

There was a mistake in the sentence on page 14. It has been corrected. For all expertiments using ROUGE as evaluation tool, the AMT summaries were compared only to the model ones. We did not perform a direct comparison between AMT summaries and Wikipedia summaries. However, we did compare the results of the Wikipedia summaries with respect to the model ones. The reason why doing this (model vs. model and wikipedia vs. model) was to have an idea of a upper bound and analyze how far our results were from them.

2. **On top of p6, authors should also include a reference to the third author's work, just published at LREC [1], in which their findings about the influence of payment on quality contradict those of Mason and Watts.**

[1] Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour and Udo Kruschwitz. assessing Crowdsourcing Quality through Objective Tasks. LREC'12.

Yes, we included the reference.

3. **Related to the above, another intriguing aspect that is not discussed is how the payment value could have interfered with the results. For example, in experiment 5 the author payed .15 per HIT (which is on average the triple of the amount payed in the other experiments) and as**

a result this batch of tasks was also completed in the shortest time span (only 2 hours). Yet the time spent by turkers on individual tasks is very low, most of the tasks being completed in an unreasonable time span of 0.5 seconds!! Is this yet another proof that too high payments attract too many cheaters? A discussion and a comparison between the payment, completion time, average task solving time and the overall result quality could be very interesting in the concluding section. Also, how does this compare against one of the author's findings in [1], where a positive correlation between payment and quality was found?

[1] Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour and Udo Kruschwitz. assessing Crowdsourcing Quality through Objective Tasks. LREC'12.

We included this into the conclusion:

Finally, the results seem to indicate that the observed low reliability may be due to motivational factors, and this aspect should be studied in future work. In previous work payment was studied as one of the motivational factor for controlling the results quality. Aker et al. (2012) showed that high payments lead to better results whereas in Mason and Watts (2010) and Feng et al., (2009) it is discussed that an increase in payment attracts more spammers and as consequence it leads to low quality results. In our summary generation experiments we varied the payments in small steps from low to high however, in overall we have not manage to obtain any useful results from the MTurk workers. As discussed above the only useful results were obtained through the experiment 3. This experiment 3 differed from the other in a way that it was far easy to complete and did not require any time consuming reading task. Thus, we think that the major factor in obtaining high quality results is the level of difficulty of a task. If a task is easy and fast to perform then we think that this will positively influence the results quality. On ther other hand if the task requires a lot of time to complete and is difficult to perform this causes that the workers loose motivation on the task which then reflects the results negatively.

4. **Typos:**

All typos and spelling errors have been corrected. Moreover, the paper has been proofread in order to avoid this type of errors.

- **p2: have been analyze => has been analyzed →** *Done*
- **p4: broadly categorize => broadly categorized →** *Done*

- **p4: $60 USD => $60 (also, pleas be consisten through the document, e.g., in some places you have "0.05 US dollars"** → *Done. All the quantities have been unified into US dollars*
- **p17: not time enough => not enough time** → *Done*
- **p18: confirms us that => confirms that** → *Done*
- **p20: we broad => we broaden** → *Done*
- **p22, caption Fig.5: th euse of the "image annotation" term does not reflect the experiment itself** → *This has been changed.*
- **p24: second sentence of Section 9.2 does not make sense** → *This sentence has been removed.*
- **p25: do not obtained => did not obtain** → *Done*
- **p27: Crowdflower through AMT => AMT through Crowdflower** → *Done*
- **p27: even being aware => even if being aware** → *Done*
- **p27: higher in complexity => more complex** → *Done*
- **p28: guideles as far as types of experiments that should and should not work => guidelines such as types of experiments that might or might not work** → *Done*

Noname manuscript No.
(will be inserted by the editor)

Analyzing the Capabilities of Crowdsourcing Services for Text Summarization

Elena Lloret · Laura Plaza · Ahmet Aker

Received: date / Accepted: date

Abstract This paper presents a detailed analysis of the use of crowdsourcing services for the Text Summarization task in the context of the tourist domain. In particular, our aim is to retrieve relevant information about a place or an object pictured in an image in order to provide a short summary which will be of great help for a tourist. For tackling this task, we proposed a broad set of experiments using crowdsourcing services that could be useful as a reference for others who want to rely also on crowdsourcing.

From the analysis carried out through our experimental setup and the results obtained, we can conclude that although crowdsourcing services were not good to simply gather gold-standard summaries (i.e., from the results obtained for experiments 1, 2 and 4), the encouraging results obtained in the third and sixth experiments motivate us to strongly believe that they can be successfully employed for finding some patterns of behaviour humans have when generating summaries, and for validating and checking other tasks. Furthermore, this analysis serves as a guideline for the types of experiments that might or might not work when using crowdsourcing in the context of text summarization.

Keywords Information retrieval · Text Summarization · Crowdsourcing services · Crowdflower · Mechanical Turk

Elena Lloret
University of Alicante
Apdo. de correos, 99
E-03080 Alicante, Spain
E-mail: elloret@dlsi.ua.es

Laura Plaza
Universidad Complutense de Madrid
C/ Prof. José García Santesmases s/n
28040 Madrid, Spain
E-mail: lplazam@fdi.ucm.es

Ahmet Aker
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP, UK
E-mail: a.aker@dcs.shef.ac.uk

1 Introduction

Natural Language Processing (NLP) tasks often require annotated data or gold-standards to carry out large experiments, as well as to evaluate the task's performance itself. However, to obtain this type of data is normally very costly and time-consuming. Some of these tasks can be done automatically if a gold-standard is available, for instance, to evaluate if a question answering system is able to provide the correct answer for a factual question; but others, in which a gold-standard is difficult to obtain, such as reading comprehension, machine translation, information retrieval or text summarization, are much more difficult to evaluate.

Crowdsourcing services, such as Amazon's Mechanical Turk (AMT)[1] or Crowdflower[2] have recently appeared as services where users (also known as requesters) can upload tasks (or HITs[3]) which are performed by anonymous people, also called workers. For each task performed, a small amount of money is given as a reward depending on the price assigned to the task. Crowdsourcing services are very appropriate for carrying out those tasks that are simple for humans, but very difficult for computers and would require a lot of time to be completed. However, the quality of the output was repeatedly brought into question (e.g. Gillick and Liu (2010)). Although these services include some quality measures for ensuring the quality of the results, and one can refuse to pay if the task is not properly done, workers can still cheat and game the system. At least, two basic issues have to be taken into consideration in an attempt to guarantee the results' quality. On the one hand, we need to ensure that workers are suitable for the task, and on the other hand, we have to check that workers do not give random answers while performing it. AMT and Crowdflower themselves, provide a qualification facility to assist quality control. Requesters can attach various qualification requirements to their task in order to force workers to meet these requirements before they are allowed to work. It is also possible to obtain the overall credibility of a worker by measuring the percentage of accepted tasks he/she has completed since joining AMT.

AMT has been used in a number of NLP tasks, such as relevance judgement (Alonso et al, 2008), image region annotation (Sorokin and Forsyth, 2008), extracting facets from documents (Dakka and Ipeirotis, 2008), or paraphrasing verbs for noun compound interpretation (Nakov, 2008). Crowdflower has also been employed for tasks such as named entity annotation (Finin et al, 2010) or corpus creation (Negri and Mehdad, 2010). However, to the best of our knowledge, none of these services has been analyzed in detail, regarding their reliability on text summarization tasks.

Therefore, the main objective of this paper is to analyze to what extent crowdsourcing services are useful for text summarization, in particular, for the task of writing human summaries that may be use as gold-standard or model summaries for evaluation. To this end, we aim to assess the reliability of AMT through CrowdFlower workers for identifying relevant information that can be used as part of a summary. If crowdsourcing proves reliable and renders model summaries of sufficient quality, it could be used to replace the very costly process of generating human (non AMT workers) written model summaries. We set up different experiments addressed to retrieve information

[1] http://www.mturk.com

[2] Crowdflower (http://crowdflower.com/) is a crowdsourcing service built on top of AMT which allows non-US citizens to run tasks on AMT.

[3] Human Intelligence Tasks.

(e.g., sentences or nuggets) worth including in a model summary. We focus on the tourist domain, and our aim is to obtain summaries that convey useful and interesting information about a specific place or object (e.g., "river Thames", "Eiffel Tower") that is shown in an image. In the first experiment we show workers an image together with the name of the object it refers to, and ten related documents retrieved from the Web. The workers are asked to select the 10 most relevant sentences to build a summary that provides useful information about the place in the image for a potential tourist. In the second experiment, we repeat the same task, but include trap sentences in the documents, in order to account for the reliability of the summaries provided in the former experiment, and to analyze whether the workers tend to cheat or not. Asking workers to select sentences may lead to some agreement problems in the information selected. However, the task of summary generation can be broadly defined, if we ask for the kind of information users are interested in about a specific place shown in an image. In light of this, we set up the experiment 3, where workers are required to write ten questions about which information they would like to know about a place. Further on, we analyze those questions, and use the most frequent ones, to ask the workers to select the sentences in the documents that contain such information (experiment 4). In the fifth experiment, we repeat the second experiment with different quality control mechanisms to ensure the commitment and skills of the workers, and evaluate their effectiveness. Finally, the last experiment (experiment 6) aims at validating the summaries generated in experiment 5 by other AMT annotators. By doing so, we can analyze whether the refinement of the task and the addition of quality control mechanisms and acceptance requirements leads to better summaries.

Results show that crowdsourcing services may be useful when simple tasks are defined. However, despite this fact, we cannot rely on the results directly, since they need to be checked afterwards in order to account for their quality. In this sense, one limitation encountered is the time spent for validating the results, sometimes taking longer than expected. Although from our experience, some of the experiments did not work as we expected, this analysis serves as a guideline of what might and might not work, when using crowdsourcing for text summarization.

The paper is structured as follows: the related work on different tasks performed with crowdsourcing services, as well as its benefits and limitations for NLP are introduced in Section 2. The description of the corpus and the crowdsourcing used and the suggested experiments is provided in Section 3. Furthermore, an individual section is devoted to each of the proposed experiments, where a description of the task as well as a detailed analysis of the results are provided. This comprises Sections 4, 5, 6, 7, 8, and 9. Finally, a general conclusion and some orientations for future work are presented in Section 10. As an appendix, we provide the definition of a complete HIT in Section 11.

2 Related Work

In this section, we describe previous literature concerning the use of crowdsourcing services for NLP. Therefore, we first explain the general use of such services. Then, we describe different mechanisms that have been proposed to ensure the quality of the results obtained. Finally, we focus on different approaches that have used AMT or

Crowdflower[4] for the specific task of text summarization, and we highlight how our work differs from them.

2.1 Crowdsourcing for Natural Language Processing

The use of crowdsourcing services for NLP tasks can be broadly categorized into two categories depending on its purpose. On the one hand, AMT is often used for data annotation (Sorokin and Forsyth, 2008; Hsueh et al, 2009), whereas on the other hand, it is also used for assessing the quality of a specific task, such as automatic question generation (Heilman and Smith, 2010), machine translation (Callison-Burch, 2009), Wikipedia articles' quality (Kittur et al, 2008) or assessing summary length for improving document search results (Kaisser et al, 2008).

Regarding the first category, in Snow et al (2008), non-expert annotations for five NLP tasks are investigated: 1) affect recognition; 2) word similarity; 3) recognizing textual entailment; 4) event temporal ordering; and 5) word sense disambiguation. The objective is to use AMT to determine if non-expert labelers can provide reliable annotations, and a high agreement between workers and existing gold-standard labels is reported. For less than 26 US dollars and 250 hours of work, they obtained 21,690 annotations for several NLP tasks, which were also acceptable with respect to their quality. Also concerning data annotation, 3,861 labels for 982 images were collected also using AMT for less than 60 US dollars in Sorokin and Forsyth (2008). However, after obtaining the corresponding labels for an image, they had to check their quality employing several consistency scores between each pair of annotations. Hsueh et al (2009) used AMT for annotating sentiment in political blog snippets, obtaining an agreement of 81.0% regarding the relevance of snippets, 81.8% whether the snippet was subjective or not, and a 61.9% whether it was positive or negative.

Concerning the use of crowdsourcing services for evaluating NLP tasks, Kittur et al (2008) proposed the evaluation of Wikipedia articles. The task consisted of rating articles according to the quality of their content on a seven-point Likert scale. This scale includes several factors, such as how well the article is written, how accurate it is, or whether it is well structured or not. Then, the results were compared to the ones rated by a group of experts, obtaining a very low correlation (0.5). Due to this fact, they took into consideration that some workers could be gaming the system, and consequently they redesigned the experiment, requiring workers to answer some basic questions of the Wikipedia articles, before allowing them to rate the articles. With this requirement, the correlation with the expert ratings improved (0.66). Also related to Wikipedia, in (Heilman and Smith, 2010) AMT is employed to rate computer-generated reading comprehension questions about Wikipedia articles. Each question was evaluated on a five-point scale, with respect to four aspects: ungrammaticality, incorrect information, vagueness and awkwardness. The results obtained from AMT were satisfactory, achieving a correlation of 0.79 between an individual rating and the mean rating for a question. Out of the Wikipedia domain, AMT has had a great acceptance for evaluating machine translation tasks. In (Callison-Burch, 2009), the feasibility to perform manual evaluations of machine translation quality through AMT is shown, as well as the possibility to create multiple reference translations. Buzek

[4] Since AMT has been employed more than Crowdflower, in this section, we mainly focus on this platform.

et al (2010) carried out a study where AMT was used for paraphrasing the source text provided as input for a machine translation system. Two HITs were set up in AMT, one for obtaining paraphrases from English documents into Chinese, and another one for the verification of the generated paraphrases. Through their study they were able to obtain 4,821 paraphrases from 1,357 sentence pairs, but they had to be verified, incurring additional costs.

2.2 Quality Control Mechanisms

The use of crowdsourcing services has both advantages and disadvantages. Regarding their benefits, they can provide a fast and relative inexpensive mechanism to carry out tasks that are simple for humans but very difficult for computers and that require a lot of human effort. The price for completing each task can be specified. For a small amount of money, normally ranging from 0.01 US dollars to 0.10 US dollars, it is possible to perform specific tasks and have them completed within a short time. In contrast, some issues about the quality of the task performed by the workers arise, since some of them will be probably enrolled in a task only for the money, providing non-sense answers in order to decrease the time they spend with the task, but at the same time, increasing their rate of payment, as they can finish more tasks. Moreover, most of the workers are non-experts and they will be only able to perform simple and short tasks, incurring in higher participation costs when more complex tasks are requested (Kittur et al, 2008). As we previously said in Section 1, when using crowdsourcing services one must take into consideration that: i) the workers are suitable for the proposed task, and ii) the answers to the task are not given randomly. Regarding these facts, AMT provides several mechanisms to help ensure the quality of the results. Firstly, each HIT can be completed by multiple workers, so that requesters can rely on the majority of the results. With this aim, Sheng et al (2008) suggested obtaining multiple labels for the same data but determining also, when and for which data should be this done (selective repeated-labeling). Secondly, workers may fulfil some requirements before allowing them to complete the task. For instance, in (Heilman and Smith, 2010) workers are required that at least 95% of their previous work has been accepted. The last mechanism concerns the rejection of the work. If the requester is not happy with the work a worker performed, it is possible to reject his/her work, and consequently, workers are not paid. Le et al (2010) tested a quality control policy consisting of an initial training period for each worker, before they could perform a task. Apart from these mechanisms, some other strategies have been developed to prevent workers from cheating. In (Sorokin and Forsyth, 2008), besides collecting multiple annotations for the same image, a grading task and a gold-standard were also employed. In the former, a worker scored the annotations of several images, whereas the gold-standard was used to measure the extent the workers' annotations deviate from the good ones. In (Heilman and Smith, 2010), a semi-automated monitoring of the rating of computer-generated questions is performed in order to reject work from workers that were randomly performing the task. In (Tang and Sanderson, 2010), some "traps" were introduced in each task with the purpose of eliminating noise. An interesting alternative to model the reliability of individual workers is presented in Snow et al (2008), which has been shown to improve significantly the quality of annotations in textual entailment and event annotation. The suggested model relies on gold-standard labels, where a small amount of expert-labeled training data is used to correct the individual biases of different non-expert annotations by means of conditional probability. Mason and Watts (2010) and Feng et al (2009) investigated the impact of payments in the quality of the results. They found that increased financial incentives improved the quantity, but not the quality, of work performed by participants. It was explained that workers who were paid more were no more motivated than workers paid less. However, Aker et al (2012) show that an increase in payment also leads in increase in result quality.

The quality control issue is crucial when using this type of services. Without conducting any quality policies nor assuring the qualification of the workers, the results obtained cannot be as good as they were supposed. In this article, through different experiments, we have studied what happens when we rely on the results of AMT annotators using the default quality control mechanisms provided by crowdsourcing services, compared to the results obtained when such quality polices are designed in advance for each specific task.

2.3 Crowdsourcing for Text Summarization

With respect to text summarization, crowdsourcing services (in particular, AMT) have not been as explored as for other tasks. Although this type of platform should be an easy way to gather reference summaries for text summarization as well to evaluate them, Gillick and Liu (2010) showed the difficulty of replicating the same readability results as in TAC[5] 2009 for summaries with non-expert judges in contrast to expert ones. Quality control policies were first established, in order to assure that only workers with a 96% HIT approval could perform the task and, in addition, if the task was finished under 25 seconds, the work was rejected. Concerning the amount of money they paid, they analyzed different compensation levels, and found out that lower compensations (0.07 US dollars per HIT) obtained higher quality results, because it seemed that this compensation level attracted workers less interested in making money and more conscious of their work. As far as the results are concerned, the AMT evaluation presented high variability. Whereas TAC assessors could roughly agree on what makes a good summary, obtaining a standard deviation of 1.0, the standard deviation computed for workers' results was doubled (2.3). However, El-Haj et al (2010) showed the appropriateness of using AMT for collecting a corpus of human-generated single-document summaries from Wikipedia and newspaper articles in Arabic. These summaries were produced by extracting the most relevant sentences of the source document and not taking more than half of the source sentences. They collected 765 summaries that were used to evaluate the corresponding automatic ones produced by several existing Arabic summarization systems and using different evaluation strategies, such as the Dice coefficient, ROUGE (Lin, 2004), or AutoSummENG (Giannakopoulos et al, 2008). Other uses of AMT for text summarization can be found in (Kaisser et al, 2008), where several experiments were carried out to account for the ideal length a summary should have for the information retrieval task.

Our work differs from the previously mentioned research in the fact that we analyze the reliability of crowdsourcing services for the multi-document text summarization task, and more specifically in the context of retrieving relevant information regarding a place or an object represented by an image, pertaining to the tourist domain (e.g., "Edinburgh Castle"). We do not attempt to use these services to obtain fast, cheap, and

[5] Text Analysis Conference: http://www.nist.gov/tac/

lots of annotated data. In contrast, our objective is to carry out an in-depth analysis of the behavior of workers towards different summarization-related tasks. We will show that relying directly on the summaries workers provided, without checking them in advance can result in very noisy output. However, we will show how crowdsourcing services are very useful for determining the type of information a summary should contain.

3 Experimental Method

3.1 Corpora and Crowdsourcing Service

For all the experiments, we focus on the tourist domain, since a real application for Text Summarization could be to provide in a short fragment of text the most important details and useful information about a tourist place by just taking a picture of the object with a mobile device (e.g., opening hours, location, price of the British Museum). Therefore, as corpus we used the image collection described in Aker and Gaizauskas (2010). This collection contains 310 different images covering 60 of the 107 object types identified from Wikipedia (e.g., *church*, *park*, *castle*, etc.). For each image, there are up to four short descriptions or model summaries. The model summaries were created manually based on image descriptions taken from *VirtualTourist* and contain a minimum of 190 and a maximum of 210 words. The place shown in each image is described by 10 documents (the top ten related web-documents automatically retrieved using the Yahoo! search engine and the toponym associated with the image as query). These documents contain very diverse information: information related to the object in the image, such as information concerning the type of the object (i.e., *Angkor Wat is a temple...*), where it is located and when it was built (i.e., *The temple is located at Angkor, Cambodia, and was built in the early 12th century.*), background information (i.e., *It is dedicated to the Buddhist god Vishnu,...*), etc.; but also contain other completely unrelated and noisy information, such as nearby hotels and other tourist services, advertisement from the website that hosts the information, etc. Besides, the documents are highly redundant, and the information is often repeated across the different documents. The corpus also includes a Wikipedia-based summary taking the first 200 words of the corresponding Wikipedia entry.

As crowdsourcing service, we selected Crowdflower[6], since AMT cannot be directly accessed outside the U.S. Crowdflower allows the same functionalities as AMT, and indeed, it uses AMT workers. Different from AMT, it provides its own quality control mechanisms that can be employed to filter bad workers. To that end, one can create "gold units", which are questions that workers must correctly complete before they can participate in the tasks.

3.2 Experiments

Six different experiments are set up and performed following a logical order, the output of some of them being used for the other experiments. The first experiment aims at asking workers to produce a model summary that satisfies the information needs of

[6] http://crowdflower.com/

potential tourists, by selecting the top 10 sentences of the documents, most relevant to the image shown. The second one asks the same task, but the documents include some "trap" sentences, so that we can analyze to what extent workers are intentionally cheating when carrying out the task. The third and the fourth experiments focus more on the type of information that humans would like to know about a place. In this way, the goal of the third experiment is to ask workers to provide 10 questions about the information they would like to obtain about the place whereas the fourth asks them to select those sentences in the documents that answer a set of questions. These questions are the most frequent ones provided by workers in the third experiment. The aforementioned experiments are completed with the last two experiments by implementing different quality control policies aiming at selecting qualified and involved workers, and validating the summaries generated by other workers.

In the following sections, we explain each experiment in detail.

4 Experiment 1: Generating Informative Summaries about Tourist Places

The objective of this experiment is to evaluate the reliability of crowdsourcing services for the task of generating informative summaries related to a tourist place. What we intend is to analyze which sentences from a set of documents related to a particular tourist place shown in an image are relevant and worth retrieving in order to come up with a summary. This is a relatively difficult task, because there is not a single definite answer, and the only manner to identify malicious answers is by going through each summary individually and examining the type of information it contains. Moreover, this task has a degree of subjectivity, since two annotators can consider different sentences as relevant.

4.1 Experimental Setup

The set of images together with the names of the corresponding places, and the 10 Web documents describing each place were presented to 5 AMT annotators through Crowdflower platform, and they were asked to select the 10 sentences that best describe the place shown in each image. They were clearly advised that the end was to build a brief summary which presents interesting and useful information about the place for tourism purposes (e.g., opening hours, location, a brief historical details, price, etc.), and that we are not interested in the description of the image itself, i.e., what you can see on it. Moreover, they were warned not to include redundant sentences.

All annotators were required to have a 95% trust rating. In addition, users were asked to write a short text providing their overall impression on the adequacy of the documents to the images. We offered 0.05 US dollars for each summary.

4.2 Qualitative Analysis

Once the results were obtained, we noticed that the task was finished rather quickly, needing only 72 hours (in total) for producing the 525 summaries. In light of this, we decided to have a look at the individual completion times, to search for clues about potential task spoils. We realized that the average time for generating a summary was

8 minutes; however some summaries were generated in less than 1 minute (3%). In fact, the minimum time spent was 23 seconds. The graphic in Fig. 1 shows the number of tasks completed by the annotators in different time spans. The fact that a significant percentage of the annotators (73%) read the 10 documents related to an image and produced the summary in less than 10 minutes, made us have a negative feeling about the quality of the generated summaries, and suspect that the annotators might have paid little attention when performing the task.

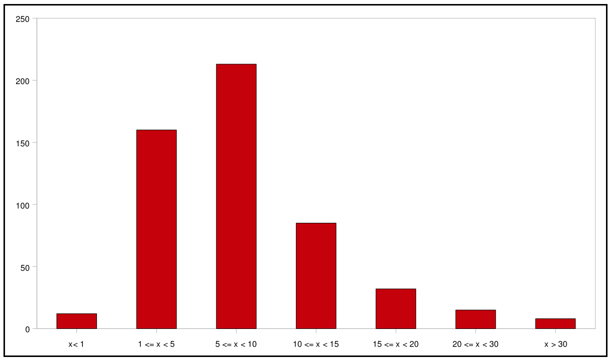

Fig. 1 Completion task time (minutes) for generating model summaries: x axis=time in minutes; y axis=number of completed tasks

In order to verify to what extent the resulting summaries are good, an analysis concerning the correlations between the annotators in their selection of the sentences was first carried out. For each image, we found that the average number of sentences that had been selected by all the annotators is 0.24, which means that most of the 5 extracts for each image have not a single sentence in common. It must be noted that, although even experts often disagree in the selection of sentences, a previous experiment performed on 25 images randomly selected showed a percentage of agreement of 3.2 between the two judges who participated in the experiment.

As it was previously anticipated, a closer look at the annotators' responses suggests some sort of random behavior. We observed, for instance, that approximately 2.5% of the extracts are composed of the first 10 sentences in the first document, even when these sentences are not at all related to the object represented in the image. Frequently, it happens that these sentences are clearly redundant or even repeated. Approximately 6% of the extracts contain the first 5 sentences in the first document, 13% present the first 3 sentences, and more than 66% of them contain the first sentence. Moreover, 17 summaries are missing, while 118 are incomplete. Surprisingly, we also found a summary containing 5 sentences that do not exist in the documents to summarize. Some summaries repeat the same sentences 2 or even 3 times. Others select the first sentence in each of the 10 documents, which often results in very poor quality summaries. To illustrated this, Table 1 shows three examples of summaries belonging to different images. Such summaries were the result of the annotators' work. As can be seen, two of these summaries only have one sentence, while the third summary presents repeated sentences, thus leading to redundant summaries (sentences 1-3 and 2-4).

Concerning the relevance of the informative content of the summaries, more than 5% of them contain advertising or announcements from the website that hosts the

Table 1 Examples of summaries generated by Crowdflower annotators

Frauenkirche, Munich: [**1**] TripAdvisor LLC is not responsible for content on external web sites.
Turaida Castle: [**1**] The castle was constructed in the Brick Gothic style in 1214 under Albert, archibishop of Riga, on the site of the destroyed wooden castle of the Livonian leader Caupo of Turaida.
Copacabana: [**1**] Copacabana beach stretches from Posto Dois (lifeguard watchtower Two) to Posto Seis (lifeguard watchtower Six). [**2**] Copacabana Beach plays host to millions of revelers during the annual New Year's Eve celebrations and, in most years, has been the official venue of the FIFA Beach Soccer World Cup. [**3**] Copacabana beach stretches from Posto Dois (lifeguard watchtower Two) to Posto Seis (lifeguard watchtower Six). [**4**] Copacabana Beach plays host to millions of revelers during the annual New Year's Eve celebrations and, in most years, has been the official venue of the FIFA Beach Soccer World Cup. [**5**] Here you will find first hand internet information about beautiful hidden brazilian paradise beaches, Samba and explosive rhythms of south America carnival brazil, barefoot dancing the new year's eve away on the warm sands of copacabana beach! [**6**] At least now you know what to expect on Copacabana beach. [**7**] With a complimentary 1-way shuttle to the airport and city center, Windsor Excelsior Hotel is across from... [**8**] Centrally located in Rio de Janeiro, this property is near Copacabana Beach and Avenida Atlantica. [**9**] During the summer international championships of beach soccer, volleyball and other sports are promoted in arenas along Copacabana Beach. [**10**] The cruisy bit is located right across from the Copacabana Palace (not shown in the picture).

information (e.g., *"TripAdvisor LLC is not responsible for content on external web sites"*). Others contain information that, even if it is related to the object in the image, do not capture important facts related to it and so should not be selected for generating the summary (e.g., *"It is hard to believe that I am the first to review Frauenkirche"*). Finally, in relation to the gramatical quality of the summaries, these are, in general, redundant and frequently present lost anaphoric and pronominal references.

Regarding the short text describing the user impression about the relevance of the information in the documents to the object in the image that was required, only 339 in the 525 annotations present such text. Within them, some just express the user opinion about the image (i.e., *"nice"* or *"I like this very much"*). Very few annotators carry out a more detailed analysis about the appropriateness of the documents for generating a summary. In this sense, we rarely find statements, such as *"Document 5 isn't relevant because it is mostly about a different temple called Ta Prohm, not Ankor Wat."* in the annotators' responses.

4.3 Quantitative Analysis using ROUGE

In order to quantitatively verify our thoughts about the poor quality of the summaries generated by the annotators, we finally evaluated them using ROUGE (Lin, 2004). ROUGE is an automatic tool for evaluating summaries that assesses the information contained in automatic summaries (called peers) by comparing them to one or more reference summaries (called models). Although we can compute several n-gram based metrics using ROUGE, for our evaluation purposes, we choose ROUGE-1, ROUGE-2 and ROUGE-SU4 metrics, since they are the most widespread in the research community. In short, ROUGE-1 and ROUGE-2 evaluate the number of identical unigram and bi-gram co-occurrences between the peer and model summaries, respectively, whereas ROUGE-SU4 allows bi-grams to have intervening word gaps no larger than four words.

We assume that human annotators will select relevant sentences from documents, and therefore, the obtained summaries will share some common vocabulary with the model summaries. Therefore, we evaluated the resulting summaries against the model summaries in the experimental corpus[7]. In addition, we compared the Wikipedia baselines to the model summaries. Note that the Wikipedia baselines are hard to surpass, since the inherent structure of Wikipedia articles tends to summarize the most important facts of the image in its first paragraph. Furthermore, we also computed ROUGE for the model summaries included in the corpus, by comparing each one against the remaining summaries of the same image. We established these results as an upper bound.

Table 2 ROGUE results for the generated summaries using crowdsourcing services, and its comparison with other human-made summaries

	Rouge-1	**Rouge-2**	**Rouge-SU4**
Model vs. Model	0.421	0.111	0.167
Wikipedia vs. Model	0.365	0.098	0.145
AMT summaries vs. Model	0.337	0.083	0.129

Table 2 shows the comparison between the human-generated summaries through crowdsourcing with respect to the model summaries. For comparison purposes, we also include the results of the model summaries compared with themselves, as well as the Wikipedia baseline compared to the model summaries. Our previous qualitative analysis of these summaries as well as our initial thoughts about their quality are confirmed by the ROUGE results obtained. As can be seen from the table, the workers annotations lead to poor quality summaries. It is worth noting that both the model summaries and the Wikipedia baseline outperform these summaries in all ROUGE metrics. A t-test was run over the results in order to account for statistical significance, and we found out that both the Wikipedia baselines and model summaries are significantly better than these summaries at a 95% confidence level.

4.4 Conclusion

Based on the results obtained, our experimental set up for using crowdsourcing services to directly build gold-standards for text summarization was not successful. ROUGE results were not as expected, since the difference between model vs. model summaries and AMT summaries vs. model ones is quite significant. One possible explanation is that building model summaries by non-expert people is really challenging, being necessary to give specific instructions to tackle it or not being so ambitious in the proposed task. However, after having a look at the summaries generated, our intuition is that most workers might not care about the task, and were more focused on the money, so they tried to complete the task as quickly as possible, as evidenced by the fact that most of the generated summaries were incomplete or included a lot of redundant or irrelevant information. Both issues will be analyzed in depth in the remaining experiments.

[7] We establish a length of 200 words for all generated summaries

5 Experiment 2: Generating Informative Summaries about Tourist Places when Trap Sentences are Included

Since the results of the first experiment were not satisfactory, in the second one we inserted some trap sentences within the documents, in order to detect whether the annotators were deliberately cheating or paying little attention to the task.

5.1 Experimental Setup

As in the previous experiment, the images together with the name of the places, and the 10 Web documents describing them were presented to 16 annotators. However, in this experiment the 10 documents contain a variable number of trap sentences. An example of trap sentence is *"The image shows the place X, but please note I am just a sentence generated by the authors and thus I am not relevant"*. They were asked to select 10 sentences from the documents, in the same conditions as experiment 1. They were advised of the existence of trap sentences that should not select. Otherwise, their work would be rejected. We offered 0.035 US dollars per each summary.

5.2 Qualitative Analysis

The users' response was even faster than in the previous experiment, spending less than 24 hours in completing the task. The average time for generating a summary was 1 minute and 20 seconds. In fact, the minimum time spent for generating a single summary was 8 seconds, and the maximum, 3 minutes and 36 seconds. The graphic in Fig. 2 shows the number of tasks completed by the annotators in different time intervals. Once again, the short time spent by the annotators to perform the task made us expect some random or neglected behavior.

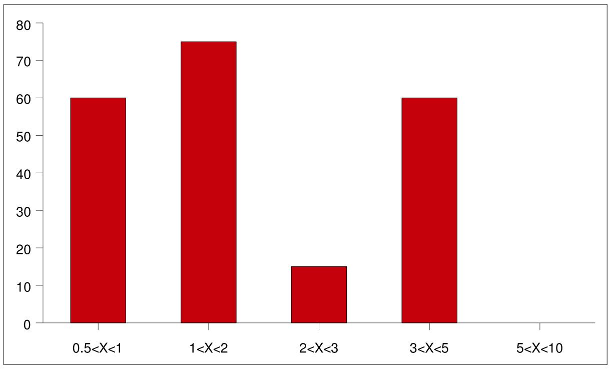

Fig. 2 Completion task time (minutes) for generating model summaries: x axis=time in minutes; y axis=number of completed tasks

Concerning correlation between the annotators, we found that the average number of sentences selected by all annotators is 0.66, which is considerably higher than in the previous experiment (0.24), but still surprisingly low.

We next analyzed the quality of the summaries. In contrast to what happened in the previous experiment, here only 3 out of the 105 summaries contain the first sentence, and only one summary contains the two first sentences, so annotators were at least selecting different sentences from the first ones. Besides, all summaries are completed, but a few of them contain an "extra" sentence. We next checked whether the summaries contain any of the trap sentences. If so, such summaries (and the workers who have generated them) should be discarded. We found that 8 of the 16 annotators had included trap sentences repeatedly in their summaries. We also found 4 of them that, instead of giving numbers representing sentences, gave letters or words (e.g., *"i,i,i,..."*, *"h,h,h,..."* or *"good,good,good,..."*). Finally, two other annotators selected the same 10 sentences for all their summaries. Therefore, we decided to ignore the summaries generated by these 14 annotators for the remaining analysis.

We next examined the summaries generated by the two remaining annotators. These summaries correspond to 30 images. We found that a good number of the summaries generated by one of the two annotators are of very poor quality, and contain very redundant and non-relevant information, including advertising from the website and hotels. This can be observed, for instance, in Table 3, where the summary for the image *Copacabana beach* is shown. This summary contains two sentences with exactly the same information (sentences 1 and 3) and other sentences with information that is clearly non-relevant to the image (sentences 5 and 8, to name a few). In contrast, the summaries generated by the other annotator are of quite good quality, as it may be seen in Table 3, where the summary for the image *Arc de Triomphe* is shown.

5.3 Quantitative Analysis using ROUGE

We next performed a quantitative evaluation of the summaries produced by the two workers using ROUGE and the same experimental set up explained in the previous experiment. Table 4 shows the comparison between these summaries with respect to the model summaries and the Wikipedia baseline. These results seem to confirm our intuition that the first annotator is selecting the sentences randomly, while the second one is paying some attention to the task. In fact, the ROUGE-1 score for the summaries produced by this last annotator are close to those of the Wikipedia baseline. In contrast, the ROUGE scores for the summaries produced by the first annotator are very poor. However, still we found that the model summaries are significantly better than all summaries from Crowdflower for all ROUGE metrics, while Wikipedia summaries are significantly better than those produced by the first annotator but equivalent to those generated by the second one for ROUGE-1 and ROUGE-SU4.

5.4 Conclusion

The results of this experiment were not satisfactory either. Even if the instructions for this task were clear and detailed, the annotators did not follow them at all, as evidenced by the fact that 8 out of the 16 annotators involved did select trap sentences. The problem, thus, does not seem to be the difficulty of the task, but the little attention paid by the workers when performing it. The annotators showed a random behaviour, and only the best annotator achieved results that were comparable to the Wikipedia results according to the ROUGE tool. In light of the results obtained in this and the

Table 3 Examples of summaries generated using crowdsourcing services

Copacabana: [**1**] As the headliner, Lenny Kravitz got to play the venue a second time, with Jorge Benjor, Macy Gray, O Rappa and Pharrell as the main opening acts, on October 2 2009, 100,000 people filled beach for a huge beach party as the IOC announced Rio would be hosting the 2016 Olympics. [**2**] As the headliner, Lenny Kravitz got to play the venue a second time, with Jorge Benjor, Macy Gray, O Rappa and Pharrell as the main opening acts, on October 2 2009, 100,000 people filled beach for a huge beach party as the IOC announced Rio would be hosting the 2016 Olympics. [**3**] Opening in September 2004, Samba City will be the great new tourist attraction of the Worlds Carnival Capitol! [**4**] But have no fear, this chili recipe with... [**5**] Taxes, fees not included for deals content. [**6**] You'll have a great time in Rio. [**7**] Key attractions in the North Zone include Maracana, one of the world's largest soccer stadiums, and Quinta da Boa Vista park, which houses the city zoo and an imperial palacea now the National Museum. [**8**] One of our. [**9**] Because Weekends are not long enough, Marriott offers the 3 night FREE when you stay 2 nights. [**10**] Copacabana also is center to the largest fireworks displays on NY's eve and its sandy beach is often used as stage for gigs, sport events and, even for sunbathing!
Arc du Triumph: [**1**] The Arc de Triomphe is a monument in Paris, France that stands in the centre of the Place Charles de Gaulle, also known as the "Place de l'etoile". [**2**] The Arc de Triomphe is one of the most famous monuments in Paris[**3**] Charles de Gaulle survived an attack upon him at the Arc de Triomphe during a parade.[**4**] After the 1963 assassination of President Kennedy, Mrs. Kennedy remembered the eternal flame at the Arc de Triomphe and requested that an eternal flame be placed next to her husband's grave at Arlington National Cemetery in Virginia.[**5**] By the early 1960s the monument had grown very blackened from coal soot and automobile exhaust, and during 1965-1966 the Arc de Triomphe was thoroughly cleaned through bleaching.[**6**] The modern-day Arc de Triomphe, surrounded by a vortex of madcap French drivers.[**7**] Looking eastwards, down the Champs Elysees, toward the Louvre, there is the Place de la Concorde, the Tuileries Gardens, and the Arc de Triomphe du Carrousel.[**8**] The Arc de Triomphe superbly tops the hill from which the Champs Elysees, the Avenue Foch, the Avenue de la Grande Armee and nine other large avenues radiate.[**9**] The Arc de Triomphe is 49m high, 45m wide and 22m deep.[**10**] All of Paris lies before your eyes from the panoramic terrace.

Table 4 ROGUE results for the generated summaries using crowdsourcing services, and its comparison with other human-made summaries

	Rouge-1	**Rouge-2**	**Rouge-SU4**
Model vs. Model	0.421	0.111	0.167
Wikipedia vs. Model	0.365	0.098	0.145
AMT-Annotator 1 vs. Model	0.260	0.042	0.081
AMT-Annotator 2 vs. Model	0.365	0.079	0.134
Averaged AMT summaries vs. Model	0.312	0.060	0.107

previous experiment, we think that it may not be feasible to use crowdsourcing services for directly building model summaries for automatic summarization. Therefore, in the following experiments, we will focus on narrowing and facilitating the task, in order to ask for more concrete information helpful for generating summaries.

6 Experiment 3: Finding the Information Users are Interested in

For the third experiment, we asked the annotators to write specific questions about what information they would like to find about a place described by an image if they were tourists visiting it (e.g., historical details, price, how to get there, etc.). It is important to note that this is not a proper summarization task. However, this information will be of great value when deciding what type of information should be selected for a summary, and consequently it can be considered a good indicator for determining sentence relevance. Moreover, it can also be employed for evaluating summaries; more specifically, for assessing to what extent the content of the summary provides the information users are interested in about a particular place.

6.1 Experimental Setup

In this experiment, we showed the annotators an image pertaining to a particular place (e.g., the *Eiffel Tower* in Paris). We also presented the annotator with the name of the place (*Eiffel Tower*) and its scene or object type (*tower*). The annotators were asked to take the role of a tourist and provide ten questions for which they would like to know the answers when they see the place shown in the image.

We showed images picturing 200 different places around the world together with their names. These places were manually selected from Wikipedia. Each image was shown to five different annotators. In total we collected 7,644 questions for 187 different places. The questions came from 184 different annotators. The expected number of questions was 10,000 (200 × 5 annotators × 10 questions). However, there are some places for which we only have questions from two or three annotators. We paid 0.05 US dollars for each image.

6.2 Qualitative Analysis

We first examined the completion time of the task, in order to account for strange patterns that could be indicative of the incorrect development of the task. The task was completed in approximately 101 hours in total. On average, each task was completed in 7 minutes. When an image of a tourist place or object together with its name is shown to a person, it is possible to think of 10 general questions about what one would like to know about it, regardless one's background knowledge, without spending a lot of time. Therefore, this duration seems quite reasonable, as this task was much more easier to perform than those proposed in experiments 1 and 2. The longest time employed was 24 minutes, while the shortest only took 24 seconds. However, in order to have a preliminary idea of the times consumed, we established different time intervals and counted the number of tasks that were completed within each interval. Results are depicted in Fig. 3. With this figure, we can have a general idea of the time employed for finishing the task of writing 10 questions. It may be seen that the completion of most tasks took between 1 and 10 minutes.

Without analyzing the content of the questions, we think that the time is quite reasonable. However, despite having reasonable time completion for most of the questions, this is not a guarantee that the annotators did their job properly. For this reason, we manually analyzed all questions in order to assess their quality. Approximately 2% of the questions were empty because not all the annotators wrote 10 questions for each object. Moreover, some questions are only related to the image itself rather than to the place shown in the image (e.g., *"when the picture is taken?", "how many flowers you found in the image?", "is there a bus in the picture?"*). Finally, there are questions

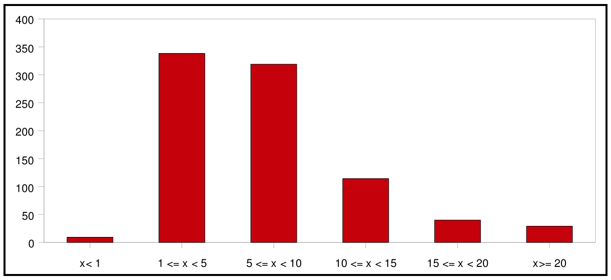

Fig. 3 Completion task time (minutes) for question generation: x axis=time in minutes; y axis=number of completed tasks

which bear no relation at all with the object in the image (e.g., *"how is the manager?"*). These questions do not address the task, which is to ask questions about the *object* shown in the image, not about the image itself or related information. Therefore, we categorized all these questions as noise which makes up 19% (1,479 out of 7,644) of the entire question set.

6.3 Quantitative Analysis

We categorized the remaining 81% of the questions (6,169 out of 7,644) by the *information type* or *attribute* the annotator was seeking with his/her question. An attribute is an abstract category grouping of similar questions. We regard two or more questions as similar if their answers refer to the same information type. For instance, we regard the questions *"where is garwood glacier?* and *"where exactly is edmonton?"* similar because both aim for answers related to the information type *location*. We name the attribute according to the information type it refers to (e.g., *location*). Table 5 shows question examples for the top five attributes (the five attributes which have the most questions).

Table 5 Top five attributes with related questions. The percentage number in brackets indicate the proportion of questions categorized by that attribute

visiting (17%)	where i can buy the ticket?, is this tower available to be visited the whole year?, when is the best time to visit?, how to get there?
location (13%)	where is garwood glacier?, where exactly is edmonton?, where it's located?
foundation-year (7%)	when was it build?, which year was this zoo opened?, when it was established?
surrounding (7%)	what are the landmarks found nearby seima palace?, what are the nearby places to visit?, what are some nearby attractioons?
features (6%)	are there any waterfalls in the park?, what does the zoo house?

6.4 Conclusion

The results obtained in this experiment are much better and accurate than in the previous experiments. One reason for this is the nature of the task, which was easier than having to produce directly extractive summaries. Regarding the questions generated, it is worth noting that half (50%) of the questions were classified within the *visiting*, *location*, *foundation-year*, *surrounding* and *features* attributes. This means that people do share ideas as to what types of information are required about a place, and the set of top five attributes captures these information types. These results are of great value when deciding what to include in the image description or summary. Therefore, we can conclude that this type of experiments are more appropriate for crowdsourcing services. In the next experiment, we verify whether this information can be used to guide and support the annotators in the summarization task.

7 Experiment 4: Finding the Answer to Specific Questions

In this experiment, we used the questions generated by the annotators in the previous experiment in order to account for the type of information a summary should contain. Therefore, the objective of this experiment is to ask the annotators to select the sentences in the documents that best answer a set of questions regarding different aspects of a place. These answers can be then put together to build up a model summary. With this experiment, we want to analyze whether there is a relationship between the information that users may be interested in by means of a set of proposed questions they have formulated in the previous experiment, and the information stated in the source documents and model summaries (its presence or its absence). Consequently, through this experiment we can study: i) whether such questions are representative of the most important information; ii) to what extent the answer to these questions appear in the documents and summaries; and finally, iii) if they can be generalized for any tourist place/object in an image.

7.1 Experimental Setup

Based on the results from experiment 3 we selected as the input for this experiment 10 questions representing the most frequent 10 types of questions from the entire set of questions. Table 6 shows such questions. This set of questions, 45 images together with the name of the object they represent, and the 10 documents describing each image were presented to 22 annotators, who were requested to select, from each set of documents, the 10 sentences which they think best answer each of the questions. As in experiment 2, the 10 documents contain some trap sentences. Also, they were told to avoid redundancy in the summaries, so in the case they found several sentences matching the same question, they could only select one. We collected a total of 312 annotations, which means an average of 6.93 annotations per image. We offered 0.04 US dollars per image.

Table 6 Questions presented to annotators

Where is it located?
How old is it?/When was it built?
Which are the dimensions of the object?
Who built it?/Who is the owner of it?
What is it famous for?/What is the history of it?
Which are the main features of it?
How can we access it?
What other attractions are nearby?
Which is the best time to visit it?
Which are best hotels nearby?

7.2 Qualitative Analysis

The experiment was completed in less than 2 hours, the averaged time for each individual task (reading the 10 documents and answering the 10 questions) being 39 seconds, which is clearly not enough time to carefully read the 10 documents and answer the 10 questions. The graphic in Fig. 4 shows the number of tasks completed within different time intervals.

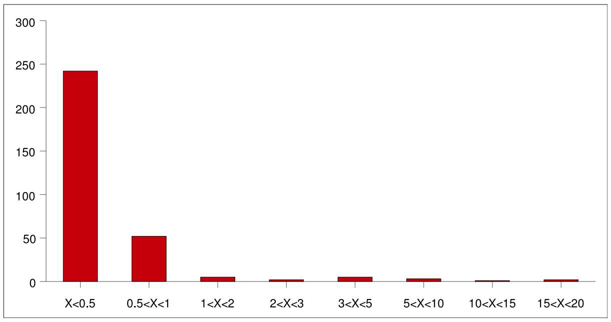

Fig. 4 Completion task time (minutes) for answering the questions: x axis=time in minutes; y axis=number of completed tasks

We next checked if the annotators had included any of the trap sentences among their answers. Although they were explicitly warned that if they included any trap their work would be rejected, we found that 20 of the 22 annotators have included them repeatedly among their answers. Moreover, they were also advised not to include repeated sentences (i.e., not to answer two or more questions with the same sentence), but we found that over 30% of the annotations present at least two repeated sentences. We also observed that approximately 15% of the annotations contain the first 5 sentences in the first document. Taking into account that all images present a trap sentence among the first 5 sentences describing them, this certainly indicates that most annotators are selecting the sentences without even having a look to the first document describing each image.

Concerning correlation between the annotators in their answers, we found that, for each task, the average number of sentences that have been selected by all annotators is

0.44. Therefore, even though this task is considerably less subjective than the previous ones, the agreement among judges remains very low.

Finally, a closer look to the responses confirms that, again, the annotators have almost certainly performed the task without paying any attention to it. It happens that some summaries repeat the same sentences even 5 times, and even worse, sometimes the redundant sentence is a trap sentence. Most of the annotations do not answer even one of the target questions. As a result, the answers can definitely not be used as model summaries for summarization evaluation. To illustrate this, Table 7 shows an example of the sentences selected by one of the annotators for the image of *Buckingham Palace* to give answers to the 10 questions in Table 6.

Table 7 Example of a summary generated through the concatenation of one annotator answers given to the 10 questions

Buckingham Palace: [**1**] Buckingham House was eventually sold by Buckingham's descendant, Sir Charles Sheffield, in 1761 to George III for 21,000 pounds. [**2**] The Belgian rooms themselves, were decorated in their present style, and named after, Prince Albert's uncle Lopold I, first King of the Belgians.[**3**] The Belgian rooms themselves, were decorated in their present style, and named after, Prince Albert's uncle Lopold I, first King of the Belgians.[**4**] The Belgian rooms themselves, were decorated in their present style, and named after, Prince Albert's uncle Lopold I, first King of the Belgians. [**5**] The Belgian rooms themselves, were decorated in their present style, and named after, Prince Albert's uncle Lopold I, first King of the Belgians. [**6**] The image potraits the place X, but please note I am just a sentence generated by the authors and thus I am not relevant. [**7**] The tradition persists of foreign. [**8**] The image contains the place X, but please note I am just a sentence generated by the authors. [**9**] The palace contains 828,818 square feet of floor space. [**10**] The image shown in the experiment features the place X, but please note I am just a sentence generated by the authors and thus I am not relevant.

7.3 Quantitative Analysis

For evaluating the results of this experiment in a quantitative manner, we did not use ROUGE, since the length of the resulting summaries was not comparable to the model ones, and consequently, comparing them using ROUGE would have distort the results and lead to incorrect interpretations. Moreover, from the qualitative analysis, we observed that the summaries were of very poor quality, and a priori, we did not know whether the source documents contain the answer to the proposed questions. Therefore, this analysis was more interesting from our point of view, since we wanted to check whether the poor quality of the summaries was due to the random behaviour of the AMT annotators (i.e., their poor commitment), or in contrast, the questions were not representative of the tourist places (i.e., their answer could not be found in the document collection). Therefore, in this experiment, we decided to analyze the number of sentences that contain the answers for the questions, both in the source documents and in the summaries, and study the existing relation.

To this end, we extracted some patterns for each question, based on its keywords, and we selected all the sentences in the documents and the corresponding summaries that match such patterns with respect to each image shown. For example, for the question *"Where is it located?"*, we took the word *"located"* as a keyword. After applying a stemming process using Porter's Stemmer, we look for sentences containing *"locat"* in this case. The stemming process will allow us to be more flexible, since the information could be expressed in the form *"X is located in . . . "*, *"X is in an ideal location in . . . "*, as well as *". . . is the location of X"*. In more general questions (e.g., *"Which are the dimensions of the object?"*), apart from looking for keywords, we also expand the keyword terms with the stem of related words (e.g., *"height"*, *"width"*, *"length"*, for the previous example).

Table 8 Average number of sentences containing the answer for the questions in the source documents and the summaries generated by the annotators (Doc= source documents; Sum=summaries)

Question	Avg. sent (Doc)	Avg. sent (Sum)
Q1: Where is it located?	10.74	0.35
Q2: How old is it?/When was it built?	20.58	1.00
Q3: Which are the dimensions of the object?	1.67	0.07
Q4: Who built it?/Who is the owner of it?	19.47	0.79
Q5: What is it famous for?/What is the history of it?	8.98	0.40
Q6: Which are the main features of it?	8.14	0.40
Q7: How can we access it?	4.16	0.09
Q8: What other attractions are nearby?	7.33	0.12
Q9: Which is the best time to vist it?	19.84	0.60
Q10: Which are best hotels nearby?	6.67	0.09

Table 8 shows the results of this analysis, both in the source documents and the summaries. As it was expected, most of the questions can be answered employing the documents, whereas those answers do not appear in the summaries, as can be deduced from their low average results. Indeed, it is worth mentioning that several equivalent answers appear in more than one document related to an image, so the annotators could easily have found the answers in the documents, if they had performed the task properly. Moreover, it is important to stress upon the fact that there are hardly documents which do not contain at least one possible answer. This depends on the place or object itself, since some specific details will not be equally provided for all of them. This happens, for instance, for the question *"What are the dimensions of the object?"*. In total, we found out that this happens only for two questions (Q3 and Q7), in the 50% and 10% of the objects, respectively. However, this does not justify the malicious behavior of the annotators. Examining in more detail the content of the summaries generated by them, we observe that only the summaries corresponding to 8 objects out of 43 contain answers to at least half of the questions (i.e., 5 questions). However, even the summaries for such objects are not sufficiently good. On the one hand, they contain a lot of redundant information, although they were told that they had to chose only one sentences in case they were several matching the same question. On the other hand, in some cases the sentence that may answer the questions is a "trap" sentence, which the annotators should avoid including, and they were told to do so. Finally, we broaden the analysis made and we report in Table 9 the number of summaries produced by the annotators that contain at least one sentence answering the corresponding question.

Table 9 Number of summaries containing at least one sentence answering the question

Question	# summaries containing the answer
Q1: Where is it located?	14
Q2: How old is it?/When was it build?	24
Q3: Which are the dimensions of the object?	4
Q4: Who built it?/Who is the owner of it?	22
Q5: What is it famous for?/What is the history of it?	15
Q6: Which are the main features of it?	15
Q7: How can we access it?	2
Q8: What other attractions are nearby?	6
Q9: Which is the best time to visit it?	21
Q10: Which are best hotels nearby?	5

7.4 Conclusion

Based on the qualitative and quantitative evaluation results, contrary to what we might have expected, the AMT annotators were showing again a random behaviour, that led to very poor results. Even though the instructions given to the annotators were clear, the task was bounded and well-defined, and the answer to most questions was easy to find in the documents, the results obtained confirmed that in most of the cases, the annotators were not paying any interest in doing the task properly. The little time spent to finish each annotation task (39 seconds on average) demonstrates that, in this case, crowdsourcing services were not appropriate for this task, which required high dedication and commitment of annotators. One of the reason why this could be happening is the definition of the task itself. As we previously stated, workers may be more comfortable with tasks that are more specific and easy to perform. However, we also think that quality control mechanisms are crucial in the proper development of the task. Therefore, in the following experiment, we will design and implement different quality control mechanisms that allow us to select committed and qualified workers.

8 Experiment 5: Generating Informative Summaries about Tourist Places using Quality Control Mechanisms

In light of the results of the previous experiments, in this experiment we set up different quality control mechanisms with the aim of ensuring that the workers selected are qualified and engaged to tackle the task.

8.1 Experimental Setup

As in experiments 1 and 2, the images together with the name of the places and the 10 web documents describing them were presented to workers. Motivated by the encouraging results of experiment 3, they were asked to think of 5 questions about the place for which they would like to know the answers and to select from the web documents the answers for each of the questions. An answer was allowed to be a full sentence or just part of a sentence[8]. The answers are used as a summary for the place. We offered 0.15 US dollars per correct annotation, and obtained 241 annotations.

In order to ensure that workers were suitable for the task, we implemented several quality control mechanisms. First, only annotators from English speaking countries[9] and with a 100% trust rating were allowed. We allowed workers from the U.S.A and India. Second, a recruitment task was added which consisted in answering the following objective and single-answer questions, which eliminates the uncertainty usually present in subjective tasks and allows us to discard those workers that are deliberately cheating: 1) *What is the name of the place for which you have entered the 5 pieces of information as answers?*, and 2) *In which country is the place located?*.

It is worth mentioning that the answer to the first question is given as part of the HIT, while the answer to the second may be easily found in the documents that the annotators were required to read. Workers who did not answer both questions correctly would not be considered for the main task. Third, as in the experiments 2 and 4, the annotators were advised of the presence of trap sentences. Finally, workers were told that they would not be paid if they did not perform the task correctly (i.e., if they did include trap sentences or redundant information as answers).

Moreover, to ensure that workers clearly understand what they are asked for, we included an example of possible summary for the place *Ararat*. In Section 11 the complete definition of this task is provided as an appendix.

8.2 Qualitative Analysis

Once again, workers' response was really fast, spending less than 2 hours in completing the task. The average time for reading the documents and selecting the answers was 1 minute and 50 seconds. Thus, the time spent is slightly higher than in previous experiments, which may be explained by the fact that the annotators had to write down answers instead of just entering sentence numbers. The graphic in Fig. 5 shows the number of tasks completed by the annotators in different time intervals.

We first examined the answers for the two recruitment questions. We found that 4 annotators did not answer them correctly. We next analyzed the quality of the main task responses. As in the second experiment, we found 2 annotators that gave random letters as answers (e.g., *hjkjkhh*, *dgoiwf*, etc.). Such annotators, besides, accepted a good number of tasks (50 and 45, respectively). We also found an annotator that repeated 5 times the name of the place in the image, and other that repeated 5 times the question *where is it located?* for all images. This time, however, we only found four summaries containing trap sentences, and only one that included repeated sentences. After removing those summaries, we got 37 summaries that, a priori, seemed to be valid and meet the desired characteristics of an informative summary. Table 10 shows two example summaries from the *Frauenkirche cathedral* and the *Holyrood Palace*.

[8] In the manual process we identified the full sentence where the part occurred and selected the entire sentence as answer.

[9] Our previous experiments have shown that the country of workers' does not have impact on the quality of the results.

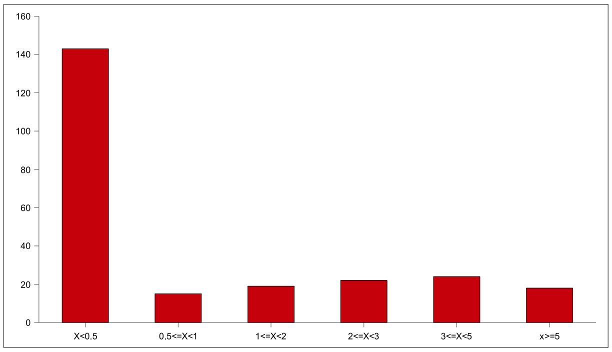

Fig. 5 Completion task time (minutes) for extracting sentences as answers to questions: x axis=time in minutes; y axis=number of completed tasks

Table 10 Examples of summaries generated using crowdsourcing services

Frauenkirche: [**1**] The Frauenkirche (full name Dom zu unserer lieben Frau, "Cathedral of Our Blessed Lady") is the cathedral of the Archbishop of Munich and Freising and is considered a symbol of the Bavarian State Capital.[**2**] The cathedral, which replaced an older romanesque church built in the 12th century, was commissioned by Duke Sigismund and erected by Jorg von Halsbach. [**3**] Construction began in 1468 and the two towers were completed in 1488. [**4**] Their design was modelled on the Dome of the Rock in Jerusalem, which in turn took a lead from late Byzantine architecture. [**5**] The Frauenkirche was constructed from red brick in the late Gothic style within only 20 years.

Holyrood Palace: [**1**] The Palace of Holyroodhouse is the official residence of the Monarch of the United Kingdom in Scotland. [**2**] The Palace stands at the bottom of the Royal Mile in Edinburgh, the opposite end to Edinburgh Castle. [**3**] It has been the site of many royal coronations and marriage ceremonies. [**4**] The palace was built around a quadrangle, situated west of the abbey cloister. [**5**] At the Palace the Queen meets and appoints the First Minister of Scotland.

8.3 Quantitative Analysis using ROUGE

We next measured the content quality of the summaries using ROUGE. Table 11 shows the comparison of the summaries generated through crowdsourcing with the model summaries and the Wikipedia baseline. Since this time the resulting summaries only contains 5 sentences or pieces of information instead of 10 sentences, the length of the model and Wikipedia summaries was truncated to 100 rather than 200 words. As a consequence, the ROUGE scores in this experiment are lower than in the previous ones. It may be seen from Table 11 that the difference between the crowdsourcing-generated summaries with respect to the Wikipedia and model summaries is less marked. However, still the model summaries are significantly better than all these summaries for all ROUGE metrics. Wikipedia summaries, in turn, are also better than crowdsourcing-generated summaries, but no significant differences exist for any of the ROUGE metrics.

Table 11 ROGUE results for the summaries generated using crowdsourcing services and its comparison with other human-made summaries

	Rouge-1	**Rouge-2**	**Rouge-SU4**
Model vs. Model	0.364	0.109	0.156
Wikipedia vs. Model	0.355	0.096	0.134
AMT summaries vs. Model	0.339	0.085	0.129

8.4 Conclusion

The results of this experiment seem to indicate that the use of control mechanisms may be of help when obtaining automatic summaries from crowdsourcing services, since they allow to quickly detect malicious user behavior. However, this control policy does not seem to guarantee the quality of the annotations, and still the results are not satisfactory. First, we have found a good number of annotators that have obviously cheated. As a consequence, more than 86% of the summaries had to be rejected. Second, the remaining annotators seem to have paid little attention to the task, as evidenced by the fact that the resulting summaries are still far from the human-made model ones.

These findings seem to confirm our intuition that the annotators are not taking the task seriously. The task seems to be excessively time-consuming and to require too much effort, so that the annotators are unwilling to perform it.

9 Experiment 6: Validating the Informative Summaries about Tourist Places using Quality Control Mechanisms

In this experiment, we assess the quality of the summaries generated in the previous experiment (Section 8). For this, we design another HIT, where the annotators rate the summaries that have been produced by other annotators. The annotators do not know where the summaries come from.

9.1 Experimental Setup

In this task, we evaluate the 37 summaries that annotators generated in the previous experiment. The HIT was designed as follows: the name of a geo-graphical place, its picture, and a description (summary) were shown to the annotators, and they were asked to rate this description within a range 1 to 5. The explanation for the ratings was also provided. A score of 1 indicated an inappropriate description, whereas the value 5 was an indication of a good description.

Finally, we evaluated the 37 summaries that annotators generated in the previous experiment. Each summary consisted of five sentences, and was evaluated by 4 different annotators. Since we consider that this task was easier to perform than the previous one, we paid 0.05 US dollars for each description rating.

In addition, the same quality control mechanisms as in the fifth experiment were taken into account (only annotators from English speaking countries with a 100% trust rating were allowed). At the end of each task, we also included two objective questions, in order to verify that the annotators were paying attention to the task. Specifically, these questions were: i) *What is the place name for which you have read its description?*, and ii) *In which country is the given place?*. As we have mentioned, the purpose of these questions was to try to discover neglected behaviour of the annotators, since the answer to the first question was already given in the HIT itself, and the answer to the second one was very easy to find in the summaries[10].

9.2 Qualitative Analysis

Contrary to what happened with the previous tasks, the average time each annotator spent for rating each summary was 3 minutes and 52 seconds, being higher than in previous tasks. Fig. 6 shows the time intervals, together with the number of tasks completed within them.

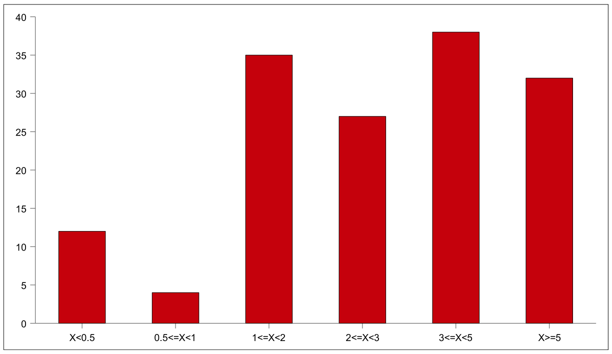

Fig. 6 Completion task time (in seconds) for summary rating: x axis=time in minutes; y axis=number of completed tasks

Taking a look at the wrong answers provided to the objective questions, we found out that 103 annotators correctly answer these questions, whereas 45 did it wrongly. The most frequent error was to copy the summary in the place of the answer. Only in one case, the annotator gave his/her personal opinion about the place as answer (*‘It is a beatiful place”*). Therefore we discarded the job of those annotators who did not perform the task properly.

9.3 Quantitative Analysis: Rating Agreement and ROUGE evaluation

Finally, we obtained 33 summaries rated, and we use them to analyze the agreement between the annotators. From the results, we realized that on average, the summaries were rated as 4, but the most frequent value assigned was 5. Most of the annotators agreed on the rating for the same summary, except for 7 summaries (i.e., 21% of disagreement), where the results highly varied depending on which annotator rated it. Table 12 shows the specific places for the summaries which did not obtained any agreement at all in their ratings. Due to the incorrect performance of the task, we did

[10] For summaries which did not contain this information we included to the end of it a sentence providing the country information.

not obtain the same number of ratings for each summary, as we initially expected (i.e., each summary rated by 4 annotators).

Table 12 Places where the annotators do not agree on the rating assigned to their summaries, together with the ratings given.

Place	Ratings
Lisbon Cathedral	1 - 5 - 3 - 4
Park of the Nations	3 - 2 - 4
Metropolitan Museum of Art	4 -2
Nottingham Castle	1 - 4 - 5
Waikiki Beach	1 - 3 - 5
Wencelas Square	5 - 2

Therefore, we did not take those summaries into account. For the remaining ones, we obtained a 79% of agreement. Then, we compute the different degrees of agreement for these summaries. Table 13 reports these values. As it can be seen, in the 51% of the cases the agreement between the annotators is above 75%.

Table 13 Degrees of agreement and percentage of summaries within them.

Agreement	Summaries
33%	6.2%
50%	9.8%
67%	12%
75%	24%
80%	3%
100%	24%

Next, we decided to carry out an analysis by taking those summaries with 100% of agreement. We wanted to check to what extent they were good and therefore, we compared them to the model summaries and the Wikipedia baselines using ROUGE, as in the previous experiments. We also set the length to 100 words, since the summaries contained only 5 sentences. Table 14 shows the recall value obtained.

Table 14 ROUGE results for the best rated generated summaries using crowdsourcing, and its comparison with other human-made summaries

	Rouge-1	Rouge-2	Rouge-SU4
Model vs. Model	0.393	0.107	0.154
Wikipedia vs. Model	0.379	0.102	0.145
AMT summaries vs. Model	0.351	0.088	0.132

As it can be seen the results of this set of summaries increase with respect to the ones shown in Table 11 by 3%, which indicates that the annotators were performing the task seriously when rating summaries. Moreover, although the results for the model and the Wikipedia summaries are higher, in this case, contrary to what occurred for the previous experiments, there was no statistical significance (according to a t-test performed) between crowdsourcing-generated summaries and the models, nor the Wikipedia and the crowdsourcing-generated summaries.

9.4 Conclusion

This experiment validated the generated summaries created by the annotators. In light of the results obtained, we can conclude that these type of tasks, where the annotators have to evaluate the work of others, are useful to discard noisy results. In addition, we confirm that it is necessary to design quality control mechanisms as well, in order to distinguish the annotators who perform the task correctly from the ones who do not pay any attention to it.

The quantitative results obtained with ROUGE showed that the performance of the summaries in which the annotators had a 100% of agreement increased with respect to the results of the whole set of summaries. This result indicates that the annotators read the summaries and rated them according to their opinion about how good they were, thus showing, in this case, more commitment to the task than in the previous experiments.

10 Conclusion and Future Work

This paper carried out a in-depth analysis of crowdsourcing services in the context of the text summarization task, by employing Crowdflower through Amazon's Mechanical Turk workers. Specifically, we focused on the identification of relevant information and the generation of multi-document summaries for providing useful details about places or objects shown in images, according to the information in which the users are interested. Through the set of proposed experiments, our aim was to determine the appropriateness of non-expert human collection of model summaries. Such experiments included different tasks, addressing the generation of extractive summaries from a collection of documents first, and focusing later on the kind of information a human is interested in, and therefore a good summary should contain. Moreover, we conducted additional experiments, where we specifically established quality control mechanisms, and acceptance requirements in order to avoid malicious workers performing the tasks.

The results obtained for experiments 1, 2 and 4 were not satisfactory, leading us to further analyze what could be happening. To this end, several reasons have been found to explain why this may occur. On the one hand, some annotators did not pay any attention to the task itself, performing it randomly, or even cheating. This was confirmed when "trap" sentences were included in the documents, and the annotators were still selecting trap sentences from the model summaries, even if being aware of this fact. On the other hand, we found that some tasks are more complex than others. This means that not all annotators are equally capable of performing the task successfully, and since they are not experts, the quality of the results may be affected. However, it is important to stress the fact that, when a less time-consuming task is requested, the quality of the results improves. That is the case of the experiment 3, where, instead of asking annotators to select the top 10 relevant sentences for describing a place shown in an image (which requires to read the entire documents related to such image), we asked them to come up with 10 questions concerning the information they wanted to know about a place. From this experiment, we obtained some specific information that users were interested in, being of great value for developing text summarization systems. From experiments 5 and 6, we can also conclude that the quality control mechanisms introduced may be of help in some cases, although they are not a guarantee that the annotators will be committed to the task, as we noticed in experiment 5, where the results were not very good either. In light of our experience, the design of a task specifically addressed to check what other annotators did may be successful, and it will allow to discard low quality results. We have demonstrated this in the experiment 6. Finally, the results seem to indicate that the observed low reliability may be due to motivational factors, and this aspect should be studied in future work. In previous work payment was studied as one of the motivational factor for controlling the results quality. Aker et al (2012) showed that high payments lead to better results whereas in Mason and Watts (2010) and Feng et al (2009) it is discussed that an increase in payment attracks more spammers and as consequence it leads to low quality results. In our summary generation experiments we varied the payments in small steps from low to high however, in overall we have not manage to obtain any useful results from the MTurk workers. As discussed above the only useful results were obtained through the experiment 3. This experiment 3 differred from the other in a way that it was far easy to complete and did not require any time consuming reading task. Thus, we think that the major factor in obtaining high quality results is the level of difficulty of a task. If a task is easy and fast to perform then we think that this will positively influence the results quality. On ther other hand if the task requires a lot of time to complete and is difficult to perform this causes that the workers loose motivation on the task which then reflects the results negatively.

One disadvantage found when using crowdsourcing services is that, in all cases, it is not possible to rely directly on the workers' annotations. A process of validation is needed afterwards to ensure the proper quality of the results. To some extent this contradicts the basic idea behind using crowdsourcing services which is to provide a framework to carry out tasks that are difficult for computers but not for humans, in a rapid and cheap way. However, the reality shows that the results provided are far from ideal, and therefore, we have to spend time checking their quality, which itself can require substantial resources.

Finally, although crowdsourcing services were not good to simply gather gold-standard summaries, the encouraging results obtained in the third and sixth experiments motivate us to strongly believe that they can be successfully employed for finding some patterns of behavior humans have when generating summaries (e.g., what type of information they usually include), and for validating and checking other tasks. From our analysis, we have provided some guidelines such as types of experiments that might or might not work when using crowdsourcing in the context of text summarization.

In the future, we would like to explore some issues in more detail. In the short-term, we will focus on analyzing the patterns found, as well as discovering additional ones, with the final purpose of improving the quality of the summaries produced by automatic systems. In the medium and long-term, we would like to design a validation HIT associated to each normal HIT, in order to study whether this strategy improves the results obtained with crowdsourcing services, and decreases the time humans have to spend checking manually the results provided by the annotators.

Acknowledgements This work was supported by the EU-funded TRIPOD project (IST-FP6-045335) and by the Spanish Government through the FPU program and the projects TIN2009-14659-C03-01, TSI 020312-2009-44, and TIN2009-13391-C04-01; and by Conselleria d'Educació - Generalitat Valenciana (grant no. PROMETEO/2009/119 and grant no. ACOMP/2010/286).

11 Appendix: HIT for Experiment 5 (Generating Informative Summaries about Tourist Places using Quality Control Mechanisms)

In this appendix, we provide the complete HIT for the experiment regarding the selection of five pieces of information or sentences that best answer different questions a person would like to know about a specific place. Together with the instructions we provide an example of how to perform the task, and we also warned the annotators of the acceptance requirements for the task to be considered correctly done.

Information Selection: Instructions

You will be given a name of a geo-graphical place, a picture of it and 10 documents related to the place. Imagine you are a tourist and have got 5 questions about the place for which you would like to know the answers (i.e., what information you would ask for). Please select your answers from the given documents. The documents contain in each line a sentence. Your answer selections can contain the entire sentence or a part of it.

Finally, you will be given two questions. Please answer them. The answers might not come from the given documents.

Example

Given Place Name: **Ararat**

Fig. 7 Example of the image shown (Mount Ararat)

Your possible questions might look like:

1. What is Ararat?
2. Where it is located?

3. What is its height?
4. How many peaks does it have?
5. When was the last erruption?

Your possible answer selections might look like:

1. Ararat is a stratovolcano, formed of lava flows and pyroclastic ejecta, with no volcanic crater.
2. Mount Ararat is located in the Eastern Anatolia Region of Turkey
3. It has an elevation of 5,137 m/16,854 ft)
4. It has two peaks
5. It is not known when the last eruption of Ararat occurred

Acceptance Requirement

A. We added trap sentences into the documents. Thus your work should be genuine. In the case a trap sentence is selected as answer, the work will be rejected.
B. You should avoid redundant information while selecting the sentences. For instance, if you select two sentences which contain the same information about the place then your work will be rejected.
C. You have to go through all documents. Otherwise your work will be rejected.

References

Aker A, Gaizauskas R (2010) Model summaries for location-related images. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta

Aker A, El-Haj M, Albakour MD, Kruschwitz U (2012) Assessing crowdsourcing quality through objective tasks. In: Proceedings of LREC

Alonso O, Rose DE, Stewart B (2008) Crowdsourcing for relevance evaluation. SIGIR Forum 42(2):9–15

Buzek O, Resnik P, Bederson BB (2010) Error driven paraphrase annotation using mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk

Callison-Burch C (2009) Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp 286–295

Dakka W, Ipeirotis PG (2008) Automatic extraction of useful facet hierarchies from text databases. In: ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, pp 466–475

El-Haj M, Kruschwitz U, Fox C (2010) Using mechanicel turk to create a corpus of arabic summaries. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta

Feng D, Besana S, Zajac R (2009) Acquiring High Quality Non-expert Knowledge from On-demand Workforce. In: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Association for Computational Linguistics, Morristown, NJ, USA, People's Web '09, pp 51–56, URL http://portal.acm.org/citation.cfm?id=1699765.1699773

Finin T, Murnane W, Karandikar A, Keller N, Martineau J, Dredze M (2010) Annotating named entities in twitter data with crowdsourcing. In: CSLDAMT '10: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Morristown, NJ, USA, pp 80–88

Giannakopoulos G, Karkaletsis V, Vouros G, Stamatopoulos P (2008) Summarization system evaluation revisited: N-gram graphs. ACM Transactions on Speech and Language Processing 5(3):1–39

Gillick D, Liu Y (2010) Non-expert evaluation of summarization systems is risky. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk

Heilman M, Smith N (2010) Rating computer-generated questions with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, pp 00–00

Hsueh PY, Melville P, Sindhwani V (2009) Data quality from crowdsourcing: a study of annotation selection criteria. In: HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, pp 27–35

Kaisser M, Hearst MA, Lowe JB (2008) Improving search results quality by customizing summary lengths. In: Proceedings of ACL-08: HLT, Columbus, Ohio, pp 701–709

Kittur A, Chi EH, Suh B (2008) Crowdsourcing user studies with mechanical turk. In: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp 453–456

Le J, Edmonds A, Hester V, Biewald L (2010) Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In: Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010), Geneva, Switzerland, pp 17–20

Lin CY (2004) ROUGE: a package for automatic evaluation of summaries. In: Proceedings of ACL Text Summarization Workshop, pp 74–81

Mason W, Watts DJ (2010) Financial Incentives and the "Performance of Crowds". SIGKDD Explor Newsl 11:100–108

Nakov P (2008) Noun compound interpretation using paraphrasing verbs: Feasibility study. In: AIMSA '08: Proceedings of the 13th international conference on Artificial Intelligence, pp 103–117

Negri M, Mehdad Y (2010) Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In: CSLDAMT '10: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Morristown, NJ, USA, pp 212–216

Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. In: KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 614–622

Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, pp 254–263

Sorokin A, Forsyth D (2008) Utility data annotation with amazon mechanical turk. pp 1–8

Tang J, Sanderson M (2010) Evaluation and user preference study on spatial diversity. In: Proceedings of the 32nd European Conference on Information Retrieval (ECIR)