	Fix		Verify	
Difficulty	BudgetFix	Soylent	BudgetFix	Soylent
Easy	4.00 (1.41)	5.16 (1.53)	4.69	7.66
Medium	4.27 (1.23)	6.50 (1.40)	3.00	7.00
Hard	4.23 (1.23)	6.71 (1.62)	3.00	8.14

Table 1: The average number of tasks in the Fix and Verify phases generated by BudgetFix(2.25, 2, 3, 0.1) and Soylent with respect to sentence complexity. Values in bracket represent the average number of tasks passed on to the Verify from the Fix phase.

algorithms generally (at least for Soylent) increases with the difficulty of the sentences. Intuitively, one would expect AMT workers would pick out more 'incorrect' mistakes from long and difficult sentences than short and easy one in the Find phase, hence requiring more Fix tasks. However, we note that more difficult sentences do not necessarily generate more tasks in the Verify phase for both BudgetFix and Soylent. In that phase, since the total number of Verifies depends on the number of unique Fixes created by AMT workers, the lower the number of unique Fixes, the lower the number of Verifies. In fact, as can be seen from Table 1, this particular BudgetFix configuration chooses fewer Fixes than Soylent, out of which even fewer unique answers are passed to the Verify phase (1.41, 1.23, 1.23 for BudgetFix compared to 1.53, 1.40, and 1.62 for Soylent). This result suggests that it is not just the complexity of a crowdsourcing task that determines the cost but, crucially, on the performance of the workers (e.g., in creating few unique Fixes) at different phases of the process.

6. CONCLUSIONS

We investigated the interdependent task allocation under budget constraints in crowdsourcing systems. In particular, we consider the FFV workflow, where the goal is to maximise the accuracy level of the outcome with respect to a budget limit of task allocation. To solve this problem, we proposed BudgetFix, a novel crowdsourcing algorithm, that efficiently identifies the total number of tasks for the Find, Fix, and Verify phases. We also proved that the algorithm can guarantee that the probability of receiving an inaccurate outcome is at most $e^{-O(B)}$ where B is the budget limit. We demonstrated through real Amazon Mechanical Turk experiments that with 16% less budget, our method can achieve similar accuracy, compared to that of an existing algorithm used in Soylent. Moreover, we also showed that by reducing the budget of BudgetFix by roughly a fourth, the accuracy would only take a maximum 2-3% hit, and by spending less than half of Soylent's average budget results in a maximum sacrifice of 7% in accuracy. In summary, BudgetFix can achieve similar accuracy, compared to Soylent, but with a significantly lower spending cost. Given this, it is very useful in applications with complex workflows and low budgets.

Note that once a phase is considered as ready, BudgetFix will move to the next phase and does not return to the previous ones. This might cause inefficiency if the output of the previous phase is inaccurate, and will affect the quality of the work done within the subsequent phases. Given this, one possible way to make BudgetFix more efficient is to allow loopy behaviour. That is, we allow BudgetFix to return to previous phases, once it detects inaccuracy of the outcome within a particular phases. However, adding such behaviour changes makes our model significantly more complex, and the techniques we use within this paper are not suitable for analysing such complex models. Therefore, as future work, we aim to extend our analysis to the abovementioned model, making BudgetFix more efficient.

Acknowledgements

This work was carried out as part of the ORCHID project funded by EPSRC (EP/I011587/1).

7. REFERENCES

[1] O. Amir, D.G. Rand, and Y. Gal. Economic games on the internet: The effect of $1 stakes. *PloS one*, 7(2), 2012.

[2] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. *COLT*, pages 41–53, 2010.

[3] Amos Azaria, Yonatan Aumann, and Sarit Kraus. Automated agents for reward determination for human work in crowdsourcing applications. *Autonomous Agents and Multi-Agent Systems*, pages 1–22, 2013.

[4] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *UIST*, pages 313–322, 2010.

[5] P. Dai, C. H. Lin, Mausam, and D. S. Weld. POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence (AIJ)*, 202:52–85, 2013.

[6] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[7] C.-J. Ho and J. Wortman-Vaughan. Online task assignment in crowdsourcing markets. In *AAAI*, pages 45–51, 2012.

[8] E. Kamar, S. Hacker, and E. Horvitz. Combining human and machine intelligence in large–scale crowdsourcing. *AAMAS*, pages 467–474, 2012.

[9] D. R. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *CoRR*, abs/1110.3564, 2011.

[10] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, February 1975.

[11] C. H. Lin, Mausam, and D. S. Weld. Dynamically switching between synergistic workflows for crowdsourcing. *AAAI*, pages 87–93, 2012.

[12] G. Little, L. B. Chilton, M. Goldman, and R. C Miller. Turkit: tools for iterative tasks on mechanical turk. *KDD-HCOMP*, pages 29–30, 2009.

[13] E. Simpson, S. Roberts, I. Psorakis, A. Smith, and Chris Lintott. Bayesian combination of multiple imperfect classifiers. *NIPS*, 2011.

[14] R. Stranders, S. Ramchurn, B. Shi, and N. R. Jennings. Collabmap: Augmenting maps using the wisdom of crowds. In *HCOMP*, 2011.

[15] L. Tran-Thanh, S. Stein, A. Rogers, and N. R. Jennings. Efficient crowdsourcing of unknown experts using multi-armed bandits. In *ECAI*, pages 768–773, 2012.

[16] L. Tran-Thanh, M. Venanzi, A. Rogers, and N. R. Jennings. Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. *AAMAS*, pages 901–908, 2013.

[17] P. Welinder, Branson S., S. Belongie, and P. Perona. The multidimensional wisdom of crowds. *NIPS*, pages 2424–2432, 2010.