Acquiring Commonsense Knowledge for Sentiment Analysis using Human Computation

Marina Boia Claudiu Cristian Musat Boi Faltings
École Polytechnique Fédérale
de Lausanne
Switzerland
firstname.lastname@epfl.ch

ABSTRACT

Many Artificial Intelligence tasks need commonsense knowledge. Extracting this knowledge with statistical methods would require huge amounts of data, so human computation offers a better alternative. We acquire contextual knowledge for sentiment analysis by asking workers to indicate the contexts that influence the polarities of sentiment words. The increased complexity of the task causes some workers to give superficial answers. To increase motivation, we make the task more engaging by packaging it as a game. With the knowledge compiled from only a small set of answers, we already halve the gap between machine and human performance. This proves the strong potential of human computation for acquiring commonsense knowledge.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human information processing, human factors*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Design; Human Factors; Experimentation; Performance

Keywords

Human Computation; Games; Sentiment Analysis; Context

1. INTRODUCTION

Many Artificial Intelligence tasks need commonsense knowledge about the world. We consider sentiment analysis - a task that requires knowledge about the polarities of words. The most successful methods use independent words as text features, typically by adding word polarity scores from sentiment lexicons. These lexicons are either manually compiled by experts or automatically learned from corpora. The models that result give 60% to 80% accuracy, well below that of humans, who have up to 90% accuracy.

These models cannot reach human-level performance partly because word polarities are context-dependent. Some words have the same polarity in every context, but others are ambiguous and their polarities vary in different contexts: a *small room* is negative, while a *small laptop* is positive. When texts are treated as sets of independent words, context is lost. However, context is not very complex to characterize. The polarities of most words have only a few exceptions, so the size of the models could be manageable if these exceptions were identified. This is very difficult to do with statistical methods, but easy for people. Therefore, we investigate how we can obtain contextual knowledge with human computation.

We ask workers to identify the contexts that influence the polarities of words. Our task is thus more challenging than traditional polarity labeling [1, 5], and people can lose motivation. To counter this, we package the task as a game. Workers play in rounds, where in each round they increase their score by submitting answers that contain a sentiment word, a context, and a polarity. We reward workers for answers that agree with those of previous workers, thus creating the illusion of synchronous player interaction. This makes the task fun and also ensures answer quality.

From only a small set of answers, we obtain good contextual knowledge. Our context-dependent lexicon improves several established, context-independent ones, halving their deficit relative to human performance. We provide a more detailed report in [2].

2. CONTEXT DEFINITION

A *phrase* is a sentiment word, while a *context* is a word that can influence a phrase's polarity. *Unambiguous phrases* have the same polarities in every context: *excellent* is always positive. *Ambiguous phrases*, however, have context-dependent polarities: *low* is positive in the context of *price* and negative in the context of *salary*. Phrases are typically organized in sentiment lexicons. Given a phrase vocabulary P, these list the default polarities of phrases: $L = \{(phr, pol) \mid phr \in P, pol \in \{pos, neg\}\}$. This poor, context-independent representation either lists ambiguous words outside their disambiguating contexts, or excludes them altogether. Instead, we consider richer, context-dependent lexicons. Given the phrase and context vocabularies P and C, these lexicons include contexts for ambiguous phrases: $CL = \{(phr, con, pol) \mid phr \in P, con \in C, pol \in \{pos, neg\}\}$.

3. HUMAN COMPUTATION TASK

We construct context-dependent lexicons through human computation. We ask workers to analyze text snippets and to submit answers that contain three elements: a phrase, a context, and a polarity. For instance, for the text *This camera has small buttons*, workers can submit the answer $(small, buttons, neg)$.

To form useful judgements, workers need to reason more elaborately, as not all phrases are ambiguous and not all contexts disambiguate. The task thus requires cognitive engagement, and workers

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577316.

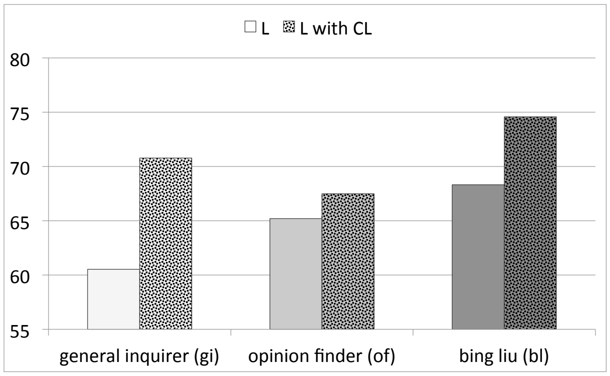

Figure 1: Lexicon performance

can lose interest. It is unclear whether relying on extrinsic motivators alone can engage workers - in previous experiments, we obtained poor results for a simple polarity annotation task where colleagues were incentivized with prizes. Therefore, to ensure workers stay motivated, we make the task fun and package it as a game.

We create the game environment with a scoring mechanism that rewards answers with common sense. An answer is commonsensical when it agrees with the common opinion of may workers. Because context can be interpreted in many ways, we cannot use a synchronous setup in which we ask pairs of workers to give identical answers. We instead choose an asynchronous setup in which we reward workers using a scoring model we compile from their activity. We reward an answer with a generous score update when it agrees with this model. We further boost the score when the answer contains an ambiguous phrase with a disambiguating context. We thus create the illusion of synchronous player interaction, which makes the task fun and ensures the quality of answers.

4. CONTEXT IMPACT

We deployed the game on Amazon Mechanical Turk and had 76 workers who submitted 6500 answers. We obtained a context-dependent lexicon CL that we tested on a product review corpus. We classified a document by obtaining a sentiment score that we thresholded at zero. We first used only a context-independent lexicon L. We scanned the document for phrases in L and updated the score using the default polarities in L. We then combined our lexicon CL with a standard lexicon L: we split the document into sentences; for each sentence, we identified the phrases that were in CL or in L; for any phrase mentioned in a context from CL, we used its polarity from CL; for all the other phrases, we used their default polarities from L.

We investigated how our context-dependent lexicon CL improved three well-established context-independent ones: General Inquirer L^{gi} [6], OpinionFinder L^{of} [7], and the lexicon of Bing Liu L^{bl} [3] (Figure 1). Alone, L^{gi} had an accuracy of 60.53%, while CL improved it by 10%. L^{of} had an accuracy of 65.20%, but CL, augmented it with 2.3%. L^{bl} produced an accuracy of 68.32%, and when combined with CL it improved by 6%. Our contextual knowledge thus successfully improved these standard lexicons.

We also analyzed how CL reduced the gap between the performance of L^{gi}, L^{of}, L^{bl} and the human performance of 83.50% [4] (Figure 2). For L^{gi}, CL reduced the gap to human performance by 44.52%. For L^{of}, we reduced the gap by 12.51%. Finally, for L^{bl} we decreased gap was by 41.23%. The contextual knowledge we gathered thus halved the deficit relative to humans.

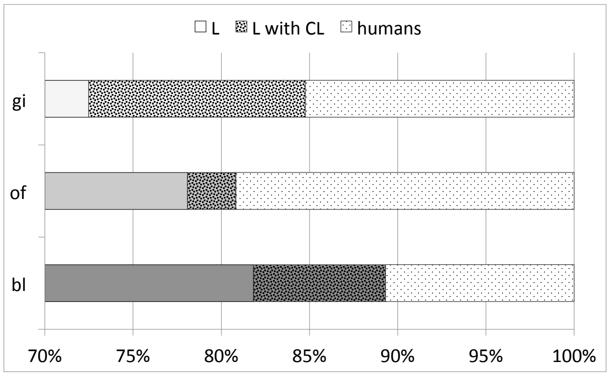

Figure 2: Lexicon performance gap relative to human performance (83.50%)

5. CONCLUSION

In many Artificial Intelligence tasks, the major difficulty is that they need commonsense knowledge. All humans share this knowledge, so it can be obtained through human computation. However, it is a big challenge to formulate tasks that are sufficiently engaging so that they produce good results. We acquired contextual knowledge for sentiment analysis. To keep workers motivated, we packaged the task as a game.

Even with a small set of answers, we obtained knowledge of good quality. We successfully improved the accuracy of three established sentiment lexicons, halving their deficit relative to human performance. We believe that a significantly larger number of answers could further improve performance.

6. REFERENCES

[1] A. Al-Subaihin, H. Al-Khalifa, and A. Al-Salman. A proposed sentiment analysis tool for modern arabic using human-based computing. In *Proceedings of the 13th International Conference on Information Integration and Web-Based Applications and Services*, pages 543–546, 2011.

[2] M. Boia, C. C. Musat, and B. Faltings. Constructing context-aware sentiment lexicons with an asynchronous game with a purpose. In *Proceedings of 15th International Conference on Intelligent Text Processing and Computational Linguistics CICLING*, 2014.

[3] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of KDD 2004*, pages 168–177, 2004.

[4] C. C. Musat and B. Faltings. A novel human computation game for critique aggregation. In *AAAI*, 2013.

[5] C. C. Musat, A. Ghasemi, and B. Faltings. Sentiment analysis using a novel human computation game. In *Proceedings of the 3rd Workshop on the People's Web Meets Natural Language Processing*, pages 1–9, 2012.

[6] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[7] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, 2005.