Infr: a video game for annotating NLI data

Justin Krasner-Karpen

1 Introduction

Having large corpuses of annotated data is critical to building functional NLP systems that require real-world knowledge. To gather this data, people have generally relied on crowdsourcing the annotation, which requires a substantial amount of money. As the part of this project done for 224N, we are building a video game to annotate data for the natural language inference task: labeling whether a pair of sentences represents an implication, a contradiction, or no inference representation. This game will task users to label both gold sentences, for which the relation is already known, as well as new sentences, both drawn from 2015's Stanford Natural Language Inference corpus, or SNLI (Bowman 2015).

2 Related Work

Previous uses of Games with a Purpose to annotate NLP data have proven quite successful. Though games such as Puzzle Racer, The Knowledge Towers, and Infection, previous studies have created data sets mapping words to images representing the same concept and associating words with related words (Vanella 2014) (Jurgens 2013). By making games that were fun and engaging to play, these projects were able to gather data more cost-effectively that through expert annotation or regular crowdsourcing, and actually ended up with more accurate labels (Vanella 2014) (Jurgens 2013). Compared to text-based Games with a Purpose, these games made use of familiar aspects of video games to draw in the player, rather than have players simply answer questions (Vanella 2014) (Jurgens 2013).

3 Methods

Moving to the natural language inference problem, it becomes challenging to turn the problem into a game. Unlike relating concepts and images like in Vanella 2014 and Jurgens 2013, NLI problems require players to read and think about large quantities of text, and require a lot of thinking power to solve.

3.1 Alternative Methods

Like in previous Games with a Purpose, we considered having a more traditional game, like a racing game, a shooter, or a platforming game, that players have to intermittently solve NLI problems to master. However, given the high cognitive load needed to resolve inferences, we didn't want players to have to think about too much in addition to the problems they were solving. So, we looked for models of successful games that involve thinking about similar problems.

3.2 Design of Infr

The success of games such as Trivia Crack prove that people can find a lot of fun in answering a succession of questions, provided that the questions are presented in an engaging context. In addition, many other successful smartphone games involve at the heart repeatedly performing a relatively simple task, yet still engage players over time with their exciting interfaces and tools like a sense of progression. The important thing to keep in mind is that NLI is a natural language task, and properly engaging with it requires thinking about the task itself. So, the challenge became thinking about a game-like smartphone interface that regularly drew people in, and that like NLI involved the theme of pairing things off with each other.

The answer we came up with was swipe dating apps. This transformed the dry task of deciding whether sentences have an entailment, neutral, or contradictory relationship into a more engaging theme of matchmaking cartoon sentences, deciding which should date, be friends, or stay away from one another. Modeling the overall structure of the game after the app Tiny Wings, we made the goal to get the most correct answers within a flexible time limit: getting answers wrong would incur a time penalty, while correct answers would give a small time bonus. Some pairs of sentences were drawn from a gold data set with known labels, while some were drawn from an unknown data set with unknown labels.

Of course, since the system has to label new data, it will say any answer to previously unseen data is correct. We chose randomly at each pair whether to give a sentence with a known or unknown answer, giving more unknown answers as the player answered more questions correclly, like in Puzzle Racer (Jurgens 2013). Unfortunately, due to time constraints we were only able to finish a text-based prototype of this game

To test the game, we had a total of 7 people unfamiliar with the NLI task play through the game, and recorded their scores as well as their responses to any questions from the unknown data set. We used SNLI's training data set as the gold data set, and its test data set as the unknown set. Since the test data set has labels, we are able to evaluate players' success in the game. In addition, we kept notes on

players' experience with the game, to later think about how to design the game to be more fun and engaging.

4 Results and Analysis

4.1 Analysis of a computer system

First, we take a look at how a computer system handles the task of analyzing NLI data. To do this, we will look at the lexicalized classifier used in the SNLI paper (Bowman 2015), and analyze the types of errors it tends to make, comparing them to how humans playing a video game might act. [1] This classifier uses 3 unlexicalized and 3 lexcalized features comparing the hypothesis to the premise.

While the classifier was not very complex, it still performed remarkably well. On the SNLI test data set, it got an accuracy of 78.2% (Bowman 2015). The classifier made a few patterns of errors, reflecting how it worked through the problem. One type of error the classifier often made was missing contradictions when the arguments of the sentence were rearranged, such as in the pair *A man wearing padded arm protectionis being bitten by a German shepherd dog/A man bit a dog* (Bowman 2015). This happened because the classifier lacked information on the syntactic structure of the sentences, and could be prevented by including such information (Bowman 2015). On the other hand, a human playing a video game isn't likely to make this sort of error. Even in a time-constrained situation, native speakers of a language should still be able to quickly pick out the topic or subject of a sentence.

In addition to struggling with the structure of the sentence overall, the classifier also misses out on connections between smaller parts. Due to the scarcity of the data set, the classifier misses out on connections even between common pairs of words like *beach/surf* and *sprinter/runner* (Bowman 2015). A human player would pick up on these differences easily, and a computer system could be adapted to handle these differences partly by working with a large dataset of Engilsh in general, not just pairs of sentences from SNLI. Using hand-labeled relationships between large numbers of words, such as by using WordNet, would help adapt to the scarcity of information. In addition, since the classifier works with individual words and bigrams, it struggles to pick up the relationships between phrases (Bowman 2015). For example, the classifier predicts a contradiction relationship for *A male is placing an order in a deli/A man buying a sandwich at a deli*, missing that *buying a sandwich* could be an example of *placing an order* (Bowman 2015). Us-

[1] We tried to run this classifier on our own machine to have a more detailed look at the errors it tended to make, but after hours of work, we were still unable to get the program to run.

ing compositional semantics to understand the meanings of entire phrases would help with this issue (Bowman 2015).

Some of these issues rely on a more fundamental issue: the classifier only knows information about sentences, and lacks knowledge about the world in general (Bowman 2015). As an example, the classifier labels *A race car driver leaps from a burning car/A race car driver escaping danger* as a contradictory pair, since it would need to know information about the world at large to know that in the context of racing cars, getting away from a burning car entails escaping danger (Bowman 2015). Handling this sort of pair with a computed system is tricky, since it moves beyond just analyzing the linguistics of the sentences. However, to approach how humans handle the task, incorporating world knowledge is necessary. One possible direction would be to look through a database of information like Wikipedia, and keep track of factual information that way.

4.2 Analysis of human data

On the unknown data set, players of Infr were only able to correctly label 63.3% of sentences, losing to the computer by 15%. There are a few possible explanations for this. For one, the players playing the game were all playing for the first time, and may not have been used to the NLI task, and to the nature of the sentences in the SNLI data set. One way to improve the success of Infr players would be to entice them to play more.

If we look at the sentences people missed, we can see patterns that inform us about how humans dealt with the task. One thing to notice is that humans' world knowledge caused them to have the opposite problem to computers: rather than lacking knowledge about the world, humans seemed to know too much about the world, leading them to overthink questions. For example, one player labelled the pair *A young infant cries while having his or her pajamas buttoned./A young baby smiles.* as neutral, while all five of the crowdsourced individuals who labelled the pair for the SNLI set called it a contradiction. A simple system might notice that *cries* and *smiles* are opposites, and quickly label the pair as contradictory. However, given more knowledge about the world and the behavior of babies, one can imagine a baby that is crying out while still smiling. In addition to mislabeling sentences as neutral, overthinking the situation could lead to different labels: one player labelled the pair *A man with long red-hair, a brown shirt and plaid pants sells fruit in a market./The man is intelligent* as contradictory rather than neutral, possibly making assumptions about the man based on his appearance or occupation. Overthinking causes humans to make mistakes that computers wouldn't, complicating the NLI task even further. Entailment should be matching human thinking, but does that mean computers should try to make the same kinds of mistakes as humans, even if

that would make them less accurate?

Another possibile explanation for people's mistakes might be that certain pairs of sentences are more difficult for humans to label than others. If this is the case, the data should show that sentences where the original crowdsourced workers who made the SNLI set should systematically disagree in similar places to where players of Infr had incorrect answers. The data is not strong enough to support this claim. Some mistakes lined up with pairs where the SNLI workers disagreed, but many mistakes happened where all five SNLI labelers agreed.

Finally, one player suggested to me that there was a difference between the difficulty of assigning different labels. For her, finding contradictions was easy, but distinguishing between entailment and neutral relations was harder. Looking at the data in general, the broader pattern seems to indicate that the neutral relation is difficult to work with for the human players overall. In every single mistake made on the unknown data set, either the true answer or the player's response was neutral. In 81% of these mistakes, the player incorrectly labeled either an entailment or a contradiction as a neutral relation. This leads us to consider a few possibilities.

First, it is possible that something inherent to the task itself makes working with the neutral relation more difficult, and any system processing the information would mislabel too many sentences as neutral. Another possibility is that the way humans think about the task may cause them to incorrectly label sentences as neutral more often than not. In either event, we may want to consider changing how we evaluate NLI systems. Instead of using the unaltered percent accurate responses, we may want to weight different kinds of errors, to punish mislabeling sentences as neutral less. Under the possibility that neutral mislabels come because of some issue inherent to the task, punishing this kind of error less would accurately reflect how the error is easier to make, and shouldn't be considered as bad. In the case that humans alone would label pairs as neutral too often, then adjusting how we evaluate the systems would accomplish a vital goal: it would train our NLP systems to make the same kinds of mistakes as people do, and ultimately make their responses more human.

4.3 Comparison

The differences between the errors the classifier makes with the task and the errors humans made on the task ultimately reflects the differences between how the two entities think about the problem. The classifier looks at all the little bits of data, and puts all of them together, adjusting its perception by a little bit each time it explores a new pattern. Humans, on the other hand, think about the problem more holistically, gathering their thoughts about the sentences as a whole before comparing them. This leads the classifier to miss the big picture, while humans get caught

up in their thoughts and miss smaller details.

5 Further Work

Ultimately, the goal of this project is to improve annotations. To do this, the first step would be to make the game more engaging and fun. A large part of this will be by implementing the graphical portions of the game. In addition, we plan to further design and tweak parts of the game to constantly increase how fun it can be.

One thing that surprised us about the game was how difficult it was. When thinking about the design, we imagined that people would be able to quickly sort through sentence pairs. When observing players though, and when playing the game ourselves, we found that labeling entailment for a sentence pair took a lot more time and thinking power than we imagined. This has led us to rethink our earlier dismissal of incorporating different video game elements into the game. Maybe instead of an increase in cognitive load, wrapping the NLI questions in a more traditional kind of game could serve as a welcome relief from the strain of thinking through the questions.

In addition, we would like to address the issue that players starting out with the game tend to have worse annotations by providing incentives within the game to continue playing over time. This could take a variety of forms: an online leaderboard where people compete over high scores, unlockable costumes, daily challenges to complete, ways to challenge your friends, and so on.

Because of the thought and the sheer amount of text needed to think about entailment, it is a challenging task to translate into a game, and we're looking forward to seeing what the game looks like as it becomes more fun.

6 Acknowledgements

Thank you to David Jurgens for his support and guidance throughout the project. Thank you to Gabor Angeli for sharing the code for the baseline model with me.

7 Bibliography

Jurgens, David, and Roberto Navigli. "Its All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation." (2013). Web. <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/viewFile/421/65>.

MacCartney, Bill. "Natural Language Inference." (2009). Web. <http://nlp.stanford.edu/~wcmac/papers/nli-diss.pdf>.

Vannella, Daniele, David Jurgens, Daniele Scarfini, Domenico Toscani and Roberto Navigli. "Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose" (2014). Web. <http://wwwusers.di.uniroma1.it/~navigli/pubs/ACL_2014_Vannellaetal.pdf>.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. *A large annotated corpus for learning natural language inference*. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. <http://nlp.stanford.edu/projects/snli/>.