Gamification for Word Sense Labeling

Noortje J. Venhuizen
University of Groningen
n.j.venhuizen@rug.nl

Valerio Basile
University of Groningen
v.basile@rug.nl

Kilian Evang
University of Groningen
k.evang@rug.nl

Johan Bos
University of Groningen
johan.bos@rug.nl

Abstract

Obtaining gold standard data for word sense disambiguation is important but costly. We show how it can be done using a "Game with a Purpose" (GWAP) called *Wordrobe*. This game consists of a large set of multiple-choice questions on word senses generated from the Groningen Meaning Bank. The players need to answer these questions, scoring points depending on the agreement with fellow players. The working assumption is that the right sense for a word can be determined by the answers given by the players. To evaluate our method, we gold-standard tagged a portion of the data that was also used in the GWAP. A comparison yielded promising results, ranging from a precision of 0.88 and recall of 0.83 for relative majority agreement, to a precision of 0.98 and recall of 0.35 for questions that were answered unanimously.

1 Introduction

One of the core aspects of semantic annotation is determining the correct sense of each content word from a set of possible senses. In NLP-related research, many models for disambiguating word senses have been proposed. Such models have been evaluated through various public evaluation campaigns, most notably SenseEval (now called SemEval), an international word sense disambiguation competition held already six times since its start in 1998 (Kilgarriff and Rosenzweig, 2000).

All disambiguation models rely on gold standard data from human annotators, but this data is time-consuming and expensive to obtain. In the context of constructing the Groningen Meaning Bank (GMB, Basile et al., 2012), a large semantically annotated corpus, we address this problem by making use of crowdsourcing. The idea of crowdsourcing is that some tasks that are difficult to solve for computers but easy for humans may be outsourced to a number of people across the globe. One of the prime crowdsourcing platforms is Amazon's Mechanical Turk, where workers get paid small amounts to complete small tasks. Mechanical Turk has already been successfully applied for the purpose of word sense disambiguation and clustering (see, e.g., Akkaya et al., 2010; Rumshisky et al., 2012). Another crowdsourcing technique, "Game with a Purpose" (GWAP), rewards contributors with entertainment rather than money. GWAPs challenge players to score high on specifically designed tasks, thereby contributing their knowledge. GWAPs were successfully pioneered in NLP by initiatives such as 'Phrase Detectives' for anaphora resolution (Chamberlain et al., 2008) and 'JeuxDeMots' for term relations (Artignan et al., 2009). We have developed an online GWAP platform for semantic annotation, called Wordrobe. In this paper we present the design and the first results of using Wordrobe for the task of word sense disambiguation.

2 Method

Wordrobe[1] is a collection of games with a purpose, each targeting a specific level of linguistic annotation. Current games include part-of-speech tagging, named entity tagging, co-reference resolution and

[1] http://www.wordrobe.org/

word sense disambiguation. The game used for word sense disambiguation is called *Senses*. Below we describe the design of Wordrobe and the data used for *Senses*.

2.1 Design of Wordrobe

Wordrobe is designed to be used by non-experts, who can use their intuitions about language to annotate linguistic phenomena, without being discouraged by technical linguistic terminology. Therefore, the games include as little instructions as possible. All games share the same structure: a multiple-choice question with a small piece of text (generally one or two sentences) in which one or more words are highlighted, depending on the type of game. For each question, players can select an answer or use the skip-button to go to the next question.

In order to encourage players to answer a lot of questions and to give good answers, they are rewarded in two ways: they can collect *drawers* and *points*. A drawer is simply a unit of a few questions – the more difficult the game, the fewer questions are in one drawer. By completing many drawers, players unlock achievements that decorate their profile page. While drawers are used to stimulate answering many questions, points are used to motivate players to play with attention. The points are calculated on the basis of two factors: the *agreement* with other players who answered the same question and the *bet* that the player put at stake. Players can place a bet reflecting the certainty about their answer. The bet is always between 10% and 100% of the points that a question is worth. The default choice is a bet of 10% and once a player adjusts the bet, this new value is remembered as the new preset value for the next question. Higher bets will result in higher gains when the answer is correct, and lower points when the answer is wrong. Since Wordrobe is designed to create gold standard annotations, the correct choice is not defined (this is exactly what we want to obtain!). Therefore, the points are calculated on the basis of the answers given by other players, as in Phrase Detectives (Chamberlain et al., 2008). The idea is that the majority rules, meaning that the choice that gets selected most by human players is probably the correct one. So, the more players agree with each other, the more points they gain. As a consequence, the score of a player is continually updated – even when the player is not playing – in order to take into account the answers provided by other players answering the same questions.

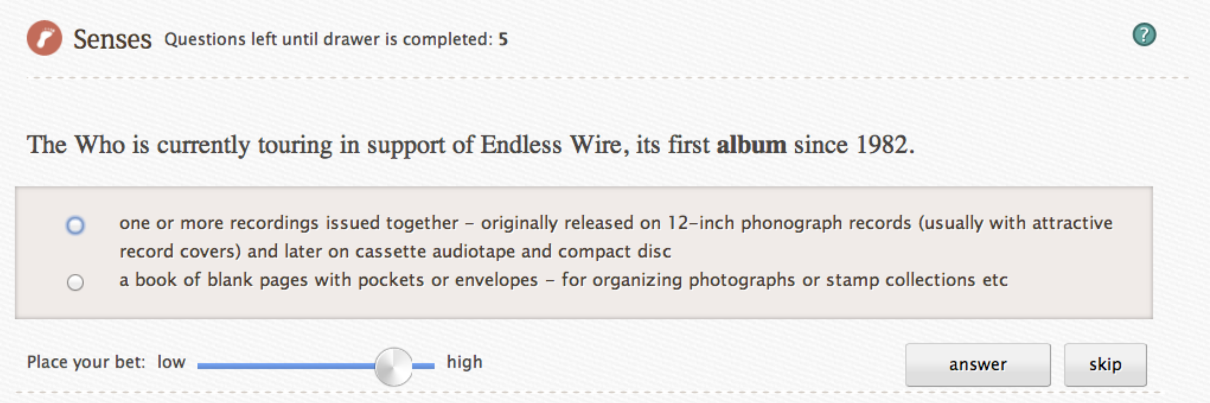
2.2 Generation of questions for the Senses game

All Wordrobe games consist of automatically generated multiple-choice questions. In the case of *Senses*, the word sense labeling game, each question consists of one or two sentences extracted from the Groningen Meaning Bank with one highlighted word for which the correct word sense in the given context must be determined. Currently, the game only focuses on nouns and verbs, but it can be easily extended to include, e.g., adjectives and adverbs. The choices for the questions are automatically generated from the word senses in WordNet (Fellbaum, 1998).

Of all the occurrences (tokens) of nouns and verbs in the GMB, 92.3% occurs in WordNet. This results in a total of 452,576 candidate questions for the *Senses* game. For the first version of Wordrobe, we selected a subset of the tokens that have at most five different senses in WordNet, such that the number of choices for each question is restricted. Figure 1 shows a screenshot of a question of *Senses*.

3 Results

The number of automatically generated questions for the first version of *Senses* was 3,121. After the first few weeks of Wordrobe going live, we had received 5,478 answers. Roughly half (1,673) of the questions received at least one answer, with an average of three answers per question. In order to test the validity of the method of using a GWAP to obtain reliable word sense annotations, we selected a subset of the questions with a reasonable response rate and created a gold standard annotation.



The Who is currently touring in support of Endless Wire, its first **album** since 1982.

○ one or more recordings issued together – originally released on 12-inch phonograph records (usually with attractive record covers) and later on cassette audiotape and compact disc

○ a book of blank pages with pockets or envelopes – for organizing photographs or stamp collections etc

Figure 1: Screenshot from Wordrobe game *Senses*.

3.1 Gold standard annotation

We created a gold standard annotation for a test set of 115 questions with exactly six answers each, which was used to evaluate the answers given by the players of Wordrobe. Four trained human annotators individually selected the correct sense for each of the target words in the test set. Fleiss's kappa was calculated to evaluate inter-annotator agreement, resulting in $\kappa = 0.79$, which is generally taken to reflect substantial agreement. Unanimity was obtained for 64% of the questions and 86% of the questions had an absolute majority vote. In a second step of evaluation, the non-unanimous answers were discussed between the annotators in order to obtain 100% agreement on all questions, the result of which was used as the gold standard annotation.

3.2 Agreement measures

Given a question and a set of player answers, we need a procedure to decide whether to accept a particular choice into our annotated corpus. One important factor is agreement: if a great majority of players agrees on the same choice, this choice is probably the correct one. Smaller majorities of players are more likely to be wrong. Another important factor is the number of answers: the more players have answered a question, the more we can presumably rely on the majority's judgement. In this work, we focus on the first factor (agreement) because the average answer rate per question is quite low throughout our data set. We tested a couple of simple agreement measures that determine whether a choice is counted as a winning answer. We measure recall and precision for each measure with respect to the gold standard.

The simplest measure accepts every choice that has a relative majority. It always accepts some choice, unless the two choices with the most answers are tied. A stricter measure ("absolute majority") accepts only the choices that were chosen by at least a certain fraction of players who answered the question, with some threshold $t \geq 0.5$. We used the values 0.5, 0.7 and 1.0 as threshold, the latter only accepting choices unanimously picked by players.

The measures described above simply choose the majority answer relative to some threshold, but fail to take into account the total number of players that answered the question and the number of possible choices for a question. These factors will become more important when we evaluate questions with a higher number of answers. We need a measure that determines whether the majority answer is chosen significantly more often than the other answers. This means that the answers should be significantly skewed towards one answer. In order to test such an effect, we can use Pearson's chi-square test, which determines the goodness-of-fit of a given distribution relative to a uniform distribution. If we take the distribution of answers over the set of possible choices, we can say that only those questions for which this distribution significantly differs from a uniform distribution ($p < 0.05$) are considered to provide an acceptable answer. Because the number of answers per question in our test set is relatively small, a significant result means that there is one choice towards which the answers accumulate. Determining which choice this is can accordingly be done using the relative-majority measure described above.

3.3 Evaluation

We evaluate the annotations obtained from Wordrobe by comparing the data of the test set (115 questions) to the gold standard. We used each of the agreement measures described above to select the answers with a high enough majority, and calculated precision (the number of correct answers with respect to the total number of selected answers), recall (the number of correct answers with respect to the total number of questions), and the corresponding F-score. The results are shown in Table 1.

Table 1: Precision and recall based on different agreement measures

Strategy	Precision	Recall	F-score
Relative majority	0.880	**0.834**	**0.857**
Absolute majority ($t = 0.5$)	0.882	0.782	0.829
Absolute majority ($t = 0.7$)	0.945	0.608	0.740
Unanimity ($t = 1$)	**0.975**	0.347	0.512
Chi-square test ($p < 0.05$)	0.923	0.521	0.666

As expected, the highest recall is obtained using the relative majority measure since this measure is the least conservative in accepting a majority choice. As the threshold for accepting a choice is set higher, recall drops and precision rises, up to a very high precision for the unanimity measure, but with a significant loss in recall. The measure based on Pearson's chi-square test is similar in being conservative; having only six answers per question in the test set, only the questions that are very skewed towards one choice give a significant result of the chi-square test.

As described above, each answer is associated with a bet between 10% and 100% of the points available for a question, which players can adjust based on how certain they are about their answer. The distribution of bets over all answers shows two significant peaks for these extremes: in 66% of the cases the maximum bet was chosen, and the default minimum bet was chosen in 12% of the cases. The main motivation for inserting the betting function was to be able to identify questions that were more difficult for players by looking for low bets. We tested the correlation between the average bet per question and the relative size of the majority (indicating agreement between players) over all questions using Pearson's product-moment correlation and found a small but significant positive effect ($r = 0.150$, $p < 0.01$). We expect that this effect will increase if more data is available.

In order to test whether questions with high average bets were easier, we repeated the evaluation, including only questions with a high average bet: $\bar{b} \geq 80\%$ (see Table 2). Recall is reduced strongly, as one would expect, but we do observe an increase in precision for all measures except unanimity. This higher precision suggests that indeed the results of the questions for which players on average place a high bet are more similar to the gold standard. However, we will need more data to confirm this point.

Table 2: Precision and recall based on different agreement measures for questions with $\bar{b} \geq 80\%$

Strategy	Precision	Recall	F-score
Relative majority	0.917	**0.478**	**0.629**
Absolute majority ($t = 0.5$)	0.930	0.461	0.616
Absolute majority ($t = 0.7$)	0.956	0.383	0.547
Unanimity ($t = 1$)	**0.961**	0.217	0.355
Chi-square test ($p < 0.05$)	0.950	0.330	0.355

4 Discussion

The goal of Wordrobe is to obtain annotations from non-expert annotators that are qualitatively close to gold standard annotations created by experts. This requires automatic techniques for filtering out low-quality answers. We evaluated the results obtained using some simple selection techniques with respect to a gold standard created by experts. We found that even with very conservative settings, optimizing for precision, we could still get a reasonably high recall (0.347). The highest precision, obtained using this most conservative measure (unanimity), was 0.975. In fact, a closer look at the data showed that there was exactly one question on which the choice unanimously picked by players differed from the gold standard annotation. This question is shown in (1).

(1) Although the last Russian **troops** left in 1994, the status of the Russian minority (some 30% of the population) remains of concern to Moscow.

a. soldiers collectively (synonyms: military personnel, soldiery)
b. a group of soldiers
c. a cavalry unit corresponding to an infantry company
d. a unit of Girl or Boy Scouts (synonyms: troop, scout troop, scout group)
e. an orderly crowd (synonyms: troop, flock)

While according to the gold standard annotation the correct answer was (1b), the six players who answered this question in the game unanimously chose (1a) as the correct answer. This example illustrates the difficulty of the task at hand very well; one could argue for the correctness of both of the possible answers. In this case, the average bet posed by the players (83%) is not helpful either in determining the difficulty of the question. This example suggests that using a more fine-grained gold standard annotation, with a ranking rather than selection of possible answers, may result in higher quality results.

Overall, the measures for calculating agreement show high numbers for precision, which were improved even more by only taking into account the questions that received a high average bet. The main drawback for this evaluation procedure is the restricted average number of answers per question. Although the recall for the unanimity measure remains at an acceptable level for the test set, this number is likely to decrease severely for questions with a higher number of answers. On the other hand, the measure based on the chi-square test is expected to become more reliable in the case of a larger dataset. In general, the evaluation measures discussed in section 3 are very basic and not robust against small datasets or unreliable annotators. With the recent uprise of crowdsourcing platforms such as Amazon's Mechanical Turk, there has been a revived interest in the task of obtaining reliable annotations from non-expert annotators. Various methods have been proposed to model annotated data such that it can be used as a gold standard (see, e.g., Carpenter, 2008; Snow et al., 2008; Beigman Klebanov and Beigman, 2009; Raykar et al., 2010). The goal of this paper was to provide a general idea of the quality of the data that can be obtained using games with a purpose. However, creation of a proper gold standard will require the collection of more data, and the use of more advanced techniques to obtain reliable annotations. As a first step towards this goal, we will make the data used in this paper available online,[2] such that interested readers can perform their own evaluation methods on this data.

5 Conclusions and future work

In this paper we described and evaluated the first results about the use of a 'Game with a Purpose' for annotating word senses. Although the amount of data obtained for each question is still relatively small (the largest amount of answers given to a reasonably sized amount of questions was 6), the results on precision and recall compared to the gold standard annotation are promising. We proposed several measures for determining the winning answer of a question, and compared them with respect to the precision and recall results. In this paper we focused on obtaining high precision scores, because the goal of the project is to obtain gold standard annotations which can be used to improve the Groningen Meaning Bank (Basile et al., 2012). Future work will focus on obtaining larger amounts of data and evaluating the annotations as part of an integration into the GMB. Moreover, this method for obtaining annotations will be applied and evaluated with respect to other linguistic phenomena, such as named entity tagging, noun-noun compound interpretation, and co-reference resolution.

[2] `http://gmb.let.rug.nl/`

Acknowledgements

We'd like to thank Bob Carpenter for comments and discussion of the Wordrobe data, and the pointers he provided to related work. Obviously we also thank all Wordrobe players (962 so far) that generated so many answers (41,541 so far) in such a short time.

References

Akkaya, C., A. Conrad, J. Wiebe, and R. Mihalcea (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 195–203. Association for Computational Linguistics.

Artignan, G., M. Hascoët, and M. Lafourcade (2009). Multiscale visual analysis of lexical networks. In *13th International Conference on Information Visualisation*, Barcelona, Spain, pp. 685–690.

Basile, V., J. Bos, K. Evang, and N. J. Venhuizen (2012). Developing a large semantically annotated corpus. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Beigman Klebanov, B. and E. Beigman (2009). From annotator agreement to noise models. *Computational Linguistics 35*(4), 495–503.

Carpenter, B. (2008). Multilevel bayesian models of categorical data annotation. Tech. report, Alias-i.

Chamberlain, J., M. Poesio, and U. Kruschwitz (2008). Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 375–380. College Publications.

Fellbaum, C. (Ed.) (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Kilgarriff, A. and J. Rosenzweig (2000). Framework and results for English SENSEVAL. *Computers and the Humanities 34*(1), 15–48.

Raykar, V., S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy (2010). Learning from crowds. *The Journal of Machine Learning Research 11*, 1297–1322.

Rumshisky, A., N. Botchan, S. Kushkuley, and J. Pustejovsky (2012). Word sense inventories by non-experts. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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