The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 82-99

Crowdsourcing Experiment Designs for Chinese Word Sense Annotation

黃資勻 Tzu-Yun Huang
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
r02142006@ntu.edu.tw

吳小涵 Hsiao-Han, Wu
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
r04142010@ntu.edu.tw

李佳臻 Chia-Chen, Lee
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
r03142008@ntu.edu.tw

李韶曼 Shao-Man, Lee
國立臺灣大學法律學系博士班
College of Law
National Taiwan University
d02a21006@ntu.edu.tw

李冠緯 Guan-Wei, Li
國立臺灣大學物理系碩士班
Master’s Program of Institute of Physics
National Taiwan University
r04222065@ntu.edu.tw

謝舒凱 Shu-Kai, Hsieh
國立臺灣大學語言所副教授
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

This paper tries to demonstrate our exploratory efforts in tackling with the “high accuracy-low quantity” problem of human word sense annotation task in Chinese, and ultimately reach the goal of automatic word sense annotation. Our proposed annotation architecture consists of explicit and implicit aspects of of crowdsourcing approach. Explicit method focuses on the general issues of crowdsourcing and made adjustments on current MTurk framework. Implicit method concentrates on the idea of Game with a Purpose (GWAP) design, which originates from a well-known video game Super Mario.

Keywords: WSD, Crowdsourcing, GWAP, Machine learning, Chinese Wordnet

1. Introduction

Sense-aware system has become central to many NLP and related intelligent systems. The core technique involved is the Word Sense Disambiguation (WSD) which can determine the proper sense of each word in varied contexts. Current WSD models rely largely on gold standard data from manual annotation that has been suffering from the problems of *high accuracy, low quantity* and *low efficiency*. This paper aims to sketch a preliminary blueprint of (word) sense annotation service by resorting to crowdsourcing (CS) approaches tailored for the Chinese WSD task.

Over the past years, crowdsourcing is an emerging collaborative way for collecting annotated corpus data and other types of language resources, with the advantages of being able to greatly increase the quantity and reduce time-cost by distribute the work to the public. Current implementations of crowdsourcing platforms include *MTurks* (e.g., Amazon Mechanical Turk; CrowdFlower), Game with a Purpose (GWAP) and *Altruistic (or volunteer-based)* crowdsourcing (e.g., Crowdcrafting). Although the explicit crowdsourcing method such as MTurks has been applied for years on several renowned platforms such as Yahoo!Answer, Quora, and so forth, several problems remain unsolved; for example, the recruitment of annotators, the annotator quality, and the design of the platforms for the recruitment. Inspired by the *CrowdTruth* project[1], we propose an internal-external adjusted framework to increase the *ground-truth* quality in the context of semantic annotation task. The explicit crowdsourcing has tackled with the main problems discovered in manual annotation; however, issue such expanses and interested-oriented bias still remain unsolved. Thus led to our second design, the implicit crowd-sourcing-game. GWAP design for annotations is not as common as the explicit approach since it is difficult to make an annotation game "interesting" and collect the required data in limited time. However, we assume that the implicit approach will become the trend by collecting data from players with greater diversities, better reflect the language user distinct, and more importantly, with low cost.

The design contributed by this paper shall be viewed as a pilot design and hope to attract relevant experts for further development. Following the introduction, Section 2 begins with a source review on English SENSEVAL, and Chinese Wordnet that we relied on, followed by a sense labelled annotation for test data and for our analysis of annotation problems in Section 3. We propose a crowdsourcing-based experiment design in Section 4, and a GWAP design in Section 5. And Section 6 concludes the paper.

2. Related Resources

SENSEVAL [1] is the international organization devoted to the sense data distribution and evaluation of Word Sense Disambiguation Systems. We use (SENSEVAL-1) sample words as our pre-selected sample. Verbs that meet the following criteria were translated into Chinese as our examples: (1) There is no homonymy, (2), the number of polysemy is between 5 and 10, and (3) the major syntactic role of the word is verb. Another resource used in this work is the Chinese Wordnet (CWN) [2], which has been developed mainly based on the English WordNet framework: synonymous lemmata are clustered as synsets, which are interconnected with various lexical semantic relations, such as antonymy, paranymy, hypernymy-hyponymy, meronymy-holonymy, etc. CWN is used as the sense

[1] http://crowdtruth.org/

inventory in this work. It is noted that in contrast with English WordNet, CWN has a higher granularity in its word meaning representation. Meaning extensions that are latent involve ‘meaning facet’, while meaning differences that are active involve ‘senses’ [17]. However, this fine-grained sense distinction is not considered for the sake of simplicity.

3. Chinese Word Sense Annotation

3.1 Data Collection and Process

Before annotation work, data collection pipeline is taken as below: word select based on Kilgarriff’s lexical sample task [3]; lemma and sense numbers confirm in Chinese Wordnet (CWN); and data collection and preprocessing. Five verbs are chosen for lexical sample task: bother (煩, fan), calculate (算, suan), float (浮,fu), invade (侵, qin), and seize (抓, zua). We translated the verbs into Chinese and remove two-word form such as 承諾 for promise, or 消耗 for consume, and look for only the ‘single character’ form with only one lemma and no more than ten senses in CWN (see Table 1).

[Table 1. Lexical sample translation, data collection and annotation assignment]

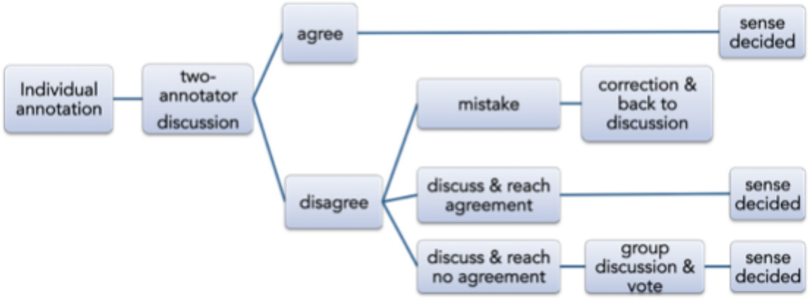
Seed word	Bother	Calculate	Float	Invade	Seize
First translations	煩, 擾	計算, 算計	漂浮	入侵, 侵入	抓, 捕
Final translation	煩 (fan)	算(suan)	浮 (fu)	侵 (qin)	抓 (zua)
Number of senses in CWN	7	10	4	6	9

3.2 Data Annotation

Five linguistic graduate students were recruited in the annotation work. Each was assigned with data collection for one verb and annotation for two verbs (Table 1.) Thus every verb was annotated by two annotators; agreement was made after every individual annotation. Data are mainly extracted from Sinica Corpus[4], and COPENS(開放語料庫) [5]. If there is no suitable concordance found in these two corpora, we search online as an alternative resource. The seed word needs to stand along as one character with one meaning, one sense.

The task was to decide and annotate the verb sense according to CWN's gloss definition. The first round was made by individual annotators without discussion with others. If there are more than one possible senses, the annotator should choose one sense and provide explanations for following discussion. In the second round, two-annotator discussion step, all sentences and tags are checked and discussed for every disagreement and ambiguity. Two annotators needed to agree to only one sense per each sentence. If not, the discussion will move on to group discussion with all team members to vote. The sense which gets most votes will be the final decision, but before the final decision, an explanation of disagreement should be provided by the annotators to other members.

There are three types of disagreements. First, mistakes from misread. Second, different interpretations of contexts. For instance, '浮' in '講到一半突然C女莫名其妙浮起來,' where '浮' can be explained as '因比重小於所在氣體而停留在該氣體中' or '在特定對象中顯現' from different perspectives. In this situation, each annotator should argue for their decision and agreed on one. Third type occur if the contextual interpretation between annotator is too different to the extent that requires all team members to discuss and vote for the final sense decision. Figure 1 shows the annotation scheme:



[Figure 1. The annotation scheme]

3.3 Annotation Problems

Three problems were found in annotation process: low-quantity, low-efficiency, and disagreements. Manual annotation is time-consuming and relatively low efficiency. And since a word may possess more than one sense and carry features from different senses in limited contexts, it often causes disagreements among annotators. To select the most suitable sense of the target word is a general but complicated issue. For human annotation, we tackle the problem by conducting cross-annotation, discussions, and vote for the best reasonable answer. But again, the time-cost is high. In order to solve the problems, we propose two possible solutions - explicit and implicit crowdsourcing designs. By outsourcing the annotation work to the public and rate annotators in advance for their credibility, the quantity may greatly increase and reduce discussion time since the one with higher score would become the agreed answer.

4. Crowdsourcing on Chinese Word Sense Tagging System

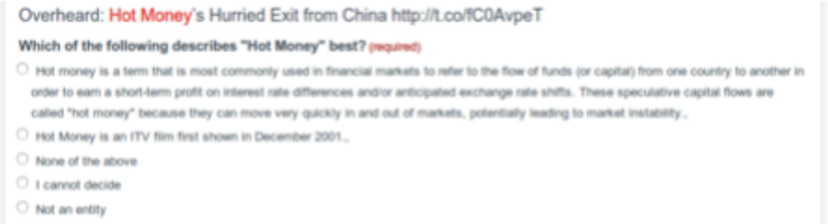
Sense annotation for Chinese WSD depends largely on manual works, which has been suffering from problems of low quantity and low efficiency. Studies before have tried to provide solutions, however, the Chinese WSD remain unsolved. The paper aims to provide solutions designed from two subtasks of the CS system.

In terms of the nature of collaboration, a CS system can be divided into two subcategories: explicit and implicit ones (Doan, Anhai, et al 2011)[6]. Two appropriate subtasks that system users can do for Chinese WSD are ‘Evaluating’: contributors assign words in context with different senses, and ’GWAP’: contributors annotate word senses through playing games in system A and contribute the game-result to system B. As an open platform for linguistic annotation, the CS system usually recruits contributors without having the ability to preview their profiles. This leads to five primary issues: the recruitment and retention of contributors, what can contributors do, how to organize the contributions, how to evaluate (Doan, Anhai, et al 2011) [4] as well as the infrastructure of system (Bontcheva, Kalina, et al 2014) [7]. Crowdsource workers can be recruited by several ways: providing payments; volunteering; by requiring; ask users to pay for the usage of system A service, then contribute to system B(crowdsourcing), such as Captcha.

As to the retention of contributors, the encouragement and retention scheme (E&R scheme) provides well-structured solutions. Systems can automatically provide instant user-gratification, display how their contributions make differences immediately.

Providing ownership is another way making users feel they own a part of the system. Previous study (Hong and Baker 2011) [8] of WSD using crowdsource approach, aggregating the inputs from contributors with majority vote. Another fact that greatly affect the results is the contributor quality, thus leads to the necessity of evaluation.The target of contributor evaluation is to prevent malicious cheating, for such problem, four solutions had been introduced by Doan in 2011. In order to manage contributors, system owner can block malicious contributors by limiting the level of contributions for individuals. We may also detect bad-intention contributions by using both manual(direct monitor) and automatic techniques(random simple question answering). Another solution is giving threat or punishment such as banning the account and public their profile. More technically, we may also create an undo system similar to Wikipedia edit page.

In order to solve previous mentioned problem, this paper provides an infrastructure of CS system for Chinese sense annotation based on the ideas of Bontcheva et al (2014)[7]. There are four main steps: first, data preprocessing; second, the creation of user interface (Figure 2 demonstrates an ideal platform for WSD crowdsourcing system (Bontcheva, Kalina et al, 2014)[7]); third, create and upload a gold unit for quality control; and last, map the judgments back to documents and aggregating them into the central database.



Overheard: Hot Money's Hurried Exit from China http://t.co/fC0AvpeT

Which of the following describes "Hot Money" best? (required)

○ Hot money is a term that is most commonly used in financial markets to refer to the flow of funds (or capital) from one country to another in order to earn a short-term profit on interest rate differences and/or anticipated exchange rate shifts. These speculative capital flows are called "hot money" because they can move very quickly in and out of markets, potentially leading to market instability.,

○ Hot Money is an ITV film first shown in December 2001.,

○ None of the above

○ I cannot decide

○ Not an entity

[Figure 2. Ideal Interface for WSD Crowdsourcing System] [7]

4.4 Design

The design of the crowdsourcing system of this paper separated into two parts, internal and external. Internally, we focused on the above-mentioned four CS-system creation steps. Externally, the main targets are the recruitment and retention of contributors and individual evaluations. Based on the consultation that CrowdFlower suggests for annotation accuracy (Hong and Baker, 2011) [8], this paper improved the infrastructure ideas (Bontcheva, Kalina et al, 2014) [7] and provides a revised framework.

4.4.1 Internal Framework

Data preparation: All pre-processed data are divided into micro-tasks with ten sentences per set to make annotation task easier. Notably, the number of senses for contributors to select from are recommended between 4 to 7, including an additional 'none of the above'.

- User interface: For better performances, instead of multiple-choice questions, users are given example sentences for each lexical item, and then asked to categorize a list of displayed sentences all at once(Hong, Baker 2011) [8]. The primary advantage is that contributors notice the difference of senses among sentences. Similar to Sinica Corpus, sentences are aligned horizontally with the target word highlighted in the page-center.

- Gold unit: In order to control quality and avoid random answers or same answers, we will set up model question and insert at least one per annotation page. A gold criterion of CrowdFlower [9] is that model questions shall be at least 20% of total questions

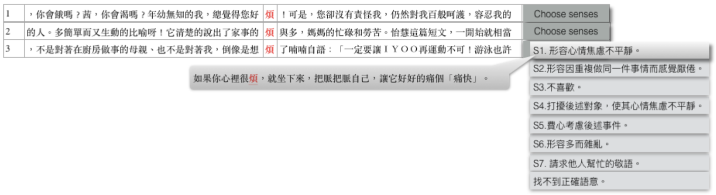
- Aggregation: Same as previous studies, this paper takes majority vote as the final result. However, for senses with equivalent score, we would recount the score of each sense based on the reliability score of individual contributors.

4.4.2 External Framework

- Recruitment and training: We provide payments to contributors; however, the payment will be retrieved if discovered cheating. The basic fee for qualified annotation is TWD 5 per set (10 sentences). Contributors with good quality will receive bonuses. Instructions will be provided in detail with explicit examples, simple terms, and avoiding jargons.

- Pre-test: Contributors are predicted to have diverse hobby of Chinese usage. By giving pre-tests on sentence understanding and meaning sensitivity before they log-in the CS system helps us control the quality and assign reliability levels of the contributors. The reliability level would effect the sense score marked by the annotator when the outcome of the annotation encountered two senses with same score and needed to be recounted.

- Crowdsourcing Micro-task: For each micro-task, contributors are required to classify sets of sentences into 4 to 7 sense categories within a single page. Once the task is finished and the results are not detected as malicious contributions, contributors will receive their rewards. Conversely, if malicious behaviors are detected, CS system will undo and remove all his or her works automatically and refuse to pay for any of his or her contributions



1	，你會餓嗎？茜，你會渴嗎？年幼無知的我，總覺得您好	煩	！可是，您卻沒有責怪我，仍然對我百般呵護，容忍我的	Choose senses
2	的人。多簡單而又生動的比喻呀！它清楚的說出了家事的	煩	與多，媽媽的忙碌和勞苦。怡慧這篇短文，一開始就相當	Choose senses
3	，不是對著在廚房做事的母親、也不是對著我，倒像是想	煩	了喃喃自語：「一定要讓ＩＹＯＯ再運動不可！游泳也許	

如果你心裡很煩，就坐下來，把脈把脈自己，讓它好好的痛個「痛快」。

[Figure 3. Revised CS User Interface for Chinese WSD Annotation]

5. **Implicit Crowdsourcing (GWAP)**

5.1 What is GWAP

GWAP, shortened for Game With a Purpose, is a sub-task of crowd-sourcing with implicit nature of collaboration, aims to solve quantity and costly issue of WSD as the explicit crow-sourcing proposed in Section 4. The definition of GWAP is: “people, as a side effect of playing, perform tasks computers are unable to perform” (Von Ahn, L., & Dabbish, L., 2008) [10]. In other words, the game developer channeled the player to work under the disguise of entertainment. The ESP Game (Google Image Labeler) is the first major success of combining game with computation task, which successfully labeled 50,000,000 images with related word. GWAP further developed in NLP field for anaphora analysis (Chamberlain et al., 2008) [11], term relations (Artignan et al., 2009) [12], semantic annotation for word sense disambiguation, known as the Wordrobe (Venhuizen, N., Basile, V., Evang, K., & Bos, J., 2013) [13], the Knowledge Towers (Vannella et al., 2014) [14], and Puzzle Racer (Jurgens, D., & Navigli, R., 2014) [15].

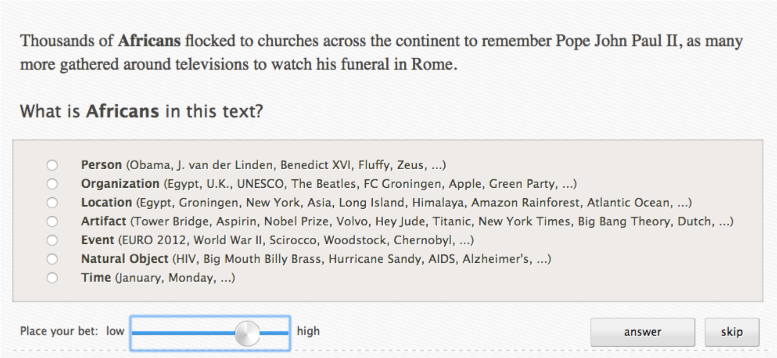
The key of a successful game is that people are willing to spend long-enough time to play, because they are ‘enjoyed’ and ‘entertained.’ And to disguise a puzzle to a game needs a well-structured design that inspires appropriate output with an enticing winning conditions and plain dos-don’ts (Von Ahn, L., & Dabbish, L., 2008) [10]. Aiming to make GWAP a universalized approach, Luis Von Ahn and Laura Dabbish addressed three templates to solve diverse computation tasks: Output-agreement games, Inversion-problem games, and Input-agreement games. And this paper is based on the output-agreement game type as design base, sharing the same initial steps and goals but with more complex winning conditions and rules. Detailed design will be elaborated in Section 5.4, followed by brief explanation of why proposing GWAP in Section 5.2, general issues and solution in Section 5.3, finally closed up by evaluation in Section 5.5.

5.2 Why GWAP

Why proposing GWAP if explicit crowd-sourcing(Section 4) can solve the quantity problem? Four major reasons are: larger amount of quantity, engaging and long-lasting; annotator diversification resulted from the game is played by layperson (Jurgens, D., & Navigli, R., 2014) [15]; better reflect native speaker instinct; and cost-down, since the game reward the player with entertainment than payment (Venhuizen, N., Basile, V., Evang, K., & Bos, J., 2013) [13].

5.3 General Issue of GWAP

Despite the advantages of GWAP, the games nowadays share some deficiencies: text-centric, randomly played, and un-controllable data gathering time. The simplest way to address text-centric WSD, is boredom, such as Wardrobe (Venhuizen, N., Basile, V., Evang, K., & Bos, J., 2013) [13], is a classic text-centric game (Figure 4). Later games developed to be more “game-centric”, hoping to create a game-like environment by transforming the senses from texts to images, such as The Knowledge Towers (Vannella et al., 2014 [14]), and Puzzle Racer(Jurgens, D., & Navigli, R., 2014[15].)



Thousands of **Africans** flocked to churches across the continent to remember Pope John Paul II, as many more gathered around televisions to watch his funeral in Rome.

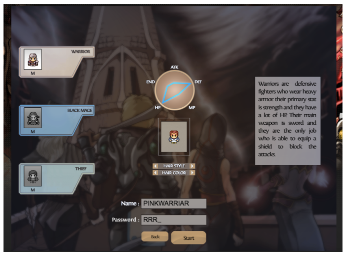
What is **Africans** in this text?

- ○ **Person** (Obama, J. van der Linden, Benedict XVI, Fluffy, Zeus, ...)
- ○ **Organization** (Egypt, U.K., UNESCO, The Beatles, FC Groningen, Apple, Green Party, ...)
- ○ **Location** (Egypt, Groningen, New York, Asia, Long Island, Himalaya, Amazon Rainforest, Atlantic Ocean, ...)
- ○ **Artifact** (Tower Bridge, Aspirin, Nobel Prize, Volvo, Hey Jude, Titanic, New York Times, Big Bang Theory, Dutch, ...)
- ○ **Event** (EURO 2012, World War II, Scirocco, Woodstock, Chernobyl, ...)
- ○ **Natural Object** (HIV, Big Mouth Billy Brass, Hurricane Sandy, AIDS, Alzheimer's, ...)
- ○ **Time** (January, Monday, ...)

Place your bet: low — high

[Figure 4. text-centric example - Wordrobe] [13]

The interface of The Knowledge Tower is a lot more game-like compare to the Wordrobe, and equipped with an import game element - my high score.

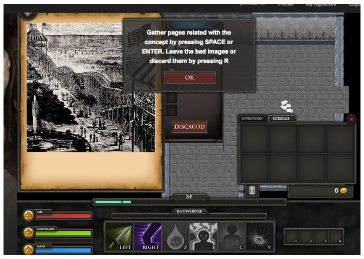


Warriors are defensive fighters who wear heavy armor. their primary stat is strength and they have a lot of HP. Their main weapon is sword and they are the only job who is able to equip a shield to block the attacks.

Name : PINKWARRIAR

Password : RRR_

[Figure 5. character selection]



Gather pages related with the concept by pressing SPACE or ENTER. Leave the bad images or discard them by pressing R

[Figure 6. Image selecting task]

The player needs to gather the images that describes the concept of the tower. The images of the senses input in the game are from an online source - Babel Net. However, we do not have a corresponding source in Chinese, it is rather difficult for the Chinese WSD game developer to replace senses with images to cut the amount of texts. How to avoid randomly played is another issue. The paper use “repeating questions” and a “player-tryout” to weight their validity. Details shall be provided in later Section.

5.4 Game for Chinese WSD

As a pioneer study of designing a game-centric GWAP for Chinese WSD, we proposed a game, “Super Chario”, named and designed after the long-lasting game “Super Mario” [16] + “Chinese.” The reason for choosing the game is to avoid players learning too many un-familiar rules and become more approachable to laymen. Since it is not yet possible to build up a WSD game based on sense images elaborated in Section 5.3, the game will focus on making text-based with challenging, entertaining, and a game-like interface.

The goal for the players is to raise an Olympia contestant, but the goal of the game is to retrieve at least 1,000 annotations per player. From average WSD annotation experiences, one may annotate 100 or more annotations per an hour. Thus we hope the game could have players to play at least 10 hours, one hour per day to reach 1,000 sentences within two weeks to control the speed of data gathering. This shall be achieved by giving “sign-in price” and “1,000 reaching price(level 50)” if they complete the challenge in 15 days.

Designs of “Super Chario” followed the game elements proposed by Von Ahn, L. et al in 2008 [10]: timed response, score keeping, player skill levels, and high score lists. The tasks needed to be completed within the time to create excitement and input-focus. Score keeping and player skill levels hope to keep the player feeling progressed. High score lists are to create an incentive for showing-off. Current architecture is specified below:

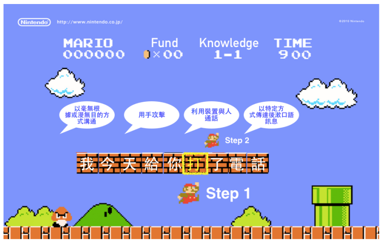
A. Initial step: After sign-up and a pre-tryout for the game, the player may choose to play by itself or with other players around you. The selection of multiple players will encounter team challenges to accomplish and create extra bonus.

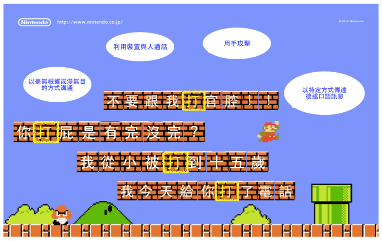
B. Winning conditions: The game is to raise an Olympia contestant by the annotations that player selects. Originally designed with 100 levels, each level contain at least 20 annotation tasks to be accomplished. Once reaching level 50 (1,000 annotations), the contestant that the player trained may write letter of challenge to battle other players to compete who's the best Chinese speaker of all time. The challenge are based on the annotation data for machine learning. One badge will be put on the cloth of the avatar every-time the player has won a battle.

C. Tasks to be accomplished

    1. Individual tasks: The task is to gain as much fund and knowledge as one can for attending the Olympia. The funding is for better geared, better food provides more energy, and change better weapons with stronger power. An individual is given three lives, if they are all used, one would not die (we do not wish to receive duplicate annotations) but need to buy a new life. Basic tasks including hitting gold words in the sentence for sense disambiguation, shoot off knowledge thieves, and grab the knowledge flag(Figure 7, source:

Super Mario). The time for response is 900 seconds per level to reduce thinking time but players may also buy time. Major way to earn funding is to touch the gold whenever you see them. Funding will also gain from expelling knowledge thieves by stepping on them or laser them with laser guns.





[Figure 7. Individual] [Figure 8. Team challenge] [Figure 9. Knowledge Monster]

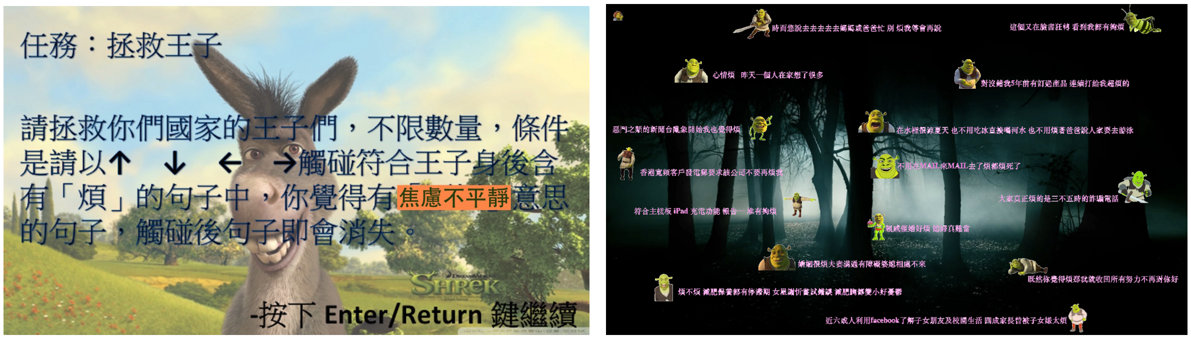
2. Team challenges: players need to drag the sentences to the possible sense and create a match. The approach hopes to encourage the player to discuss, as human annotators do if encountered disagreements(Figure 8, source: Super Mario). Aside from the annotations tasks, the team may team-up to beat the knowledge monster and earned extra funds (Figure 9, source: Super Mario).

3. Hidden tasks: Hidden tasks are in pre-selected tubes for players to earn extra funds, such as removing the sentence with different sense; or entering a sentence you think that carries the sense describe above, this may help us increase the corpus, but need to be examined later by human annotator.

4. Olympiad battle (personal machine learning): The battle is for player who has annotated more than 1,000 sentences for personal machine learning. As the player enters the Olympiad battle, they are examining their annotated results in both accuracy and recall rate, and the input questions are from previously assigned golden standard answers by trained experts.

The game-centric and data collecting time controlling is solved by using a game-like interface, multiple-tasks, “everyday sign-in price,” and “1,000 reaching price (level 50)”. Also, we could also buy the ads on Youtube or platforms to force the potential players to answer one or two questions and slowly accumulated the annotation. But how do we solve the randomly-played problem? The game borrow the weighting concept from the explicit crowd-sourcing. Upon signing up for the game, the player will be requested for a short try-out described below. Another possible approach is to repeat questions three times. The reason for repeating three times is to avoid the possibility of knowledge gain, and cause answer changed.

In order to test the weighting parameter of each player, we design a simple try-out game: “Saving Princes.” After the tryout, we would assign different titles to different players, ranging from King (Queen), Prince (Princess), Duke (Duchess), and to warriors for both gender. The game rules are as following:



任務：拯救王子

請拯救你們國家的王子們，不限數量，條件是請以↑ ↓ ← →觸碰符合王子身後含有「煩」的句子中，你覺得有焦慮不平靜意思的句子，觸碰後句子即會消失。

-按下 Enter/Return 鍵繼續

時而您說去去去去去媽媽或爸爸忙 別 煩我等會再說

這個又在臉書狂轉 看到我都有夠煩

心情煩　昨天一個人在家想了很多

對沒錯我5年前有訂過產品 連續打給我超煩的

惡鬥之類的新聞台亂象開始我也覺得煩

在水裡很涼夏天 也不用吃冰直接喝河水 也不用煩著爸爸說人家要去游泳

香港寬頻客戶發電郵要求該公司不要再煩我

不用在MAIL來MAIL去了煩都煩死了

大家真正煩的是三不五時的詐騙電話

符合主機板 iPad 充電功能 報告一堆有夠煩

親戚催婚好煩 媳婦真難當

婚姻很煩夫妻溝通有障礙婆媳相處不來

既然你覺得煩那我就收回所有努力不再對你好

煩不煩 減肥保養都有停滯期 女星諶忻嘗試錯誤 減肥胸部變小好憂鬱

近六成人利用facebook了解子女朋友及校園生活 四成家長曾被子女嫌太煩

[Figure 10. try-out game interface]

The player input their names and age. The goal of the player is to save the real princes from the dark woods. The hint of which princes are real is: find the sentences follow by the prince that fit the definition of the required sense of a particular word. For the example game attached to the paper: find the sense of “anxious (焦慮不平靜)” of “fen (煩)”. The player only needs to select the ones with the given sentences, thus the annotation numbers or meaning of the numbers provided in Table 2 shall not be relevant to the players. Sample sentences are:

[Table 2. examples of sample sentences]

Gold Standard Answer	Data
[4]	水裡很涼夏天也不用吃冰直接喝河水也不用煩著爸爸說
[1]	心情煩　昨天一個人在家想了很多
[5]	婚姻很煩夫妻溝通有障礙婆媳相處不來
[4]	對沒錯我5年前有訂過產品 連續打給我超煩的
[4]	既然你覺得煩那我就收回所有努力不再對你好

Since this is a try-out, we test only 15 sentences, however, we valued both precision and recall score of players' credibility thus we will use the F-score as the crucial criteria. If the F-score is over 70, the player would be titled as a King/Queen; over 50, the player would be Prince/Princess; over 30, the player would be Duke/Duchess; and below 30 would be all assigned as warriors. The Result shows that 2 males and 8 females with age range from 20 to 35, have played the try-out game. No players received the King/Queen title, 2 received the Prince/Princess title, 5 players were titled Duke/Duchess, and others were titled Warrior (Table 3.)

[Table 3. try-out game player result]

	Precision	Recall	F-score	Sex	Title
1	36.36	57.14	44.44	F	Duchess
2	33.33	42.86	37.50	F	Duchess
3	100.00	14.29	25.00	F	Warrior
4	50.00	57.14	53.33	M	Prince
5	25.00	14.29	18.19	F	Warrior
6	45.45	71.43	55.55	F	Princess
7	33.33	42.86	37.50	M	Duke
8	37.50	42.86	40.00	F	Duchess
9	25.00	14.29	18.19	F	Warrior
10	50.00	28.57	36.36	F	Duchess

5.5 Evaluation of GWAP

The evaluation of Super Chario may be determined by three aspects: game efficiency, player enjoyability (Von Ahn, L., & Dabbish, L., 2008), and popularity. We slightly adjust the game efficiency and player enjoyability for the purpose of evaluation, with the aid of popularity that we proposed in this paper. Game efficiency consists of "throughput" and "learning curves." Throughput is defined as the number of annotation per an hour, and the learning curves are whether a player skill strengthened overtime. A good game, in other words, is to have high throughput with learning curve slope upward. In the Super Chario, we expect the player to finish 3-4 levels per throughput, 80-100 annotations. Player enjoyability is calculated by the total amount of time played per player. The assumption is align with human intuition: we spend more time on something if we are drawn by it. Popularity is hard to measure but we might find a hint from the number of registration per day, the shape of the user growth-line since the game launched, and the ratings of the game.

Both implicit and explicit type of tasks in crowd-sourcing has their distinct advantages and disadvantages, but “correctness” is considered the major issue shared by the approaches, compared with the “golden-standard answers” annotated by trained linguistic experts. In order to measure the effectiveness, we suggest examining the annotation performances of implicit and explicit tasks by generally agreed evaluation measures in test accuracy: Precision, Recall, and F-score.

6. Conclusion

Problems witnessed in most annotation process are of annotation quantity, efficiency, and agreement. Current studies utilizing manual annotation provides only little amount of results with time-consuming and of efficiency concerns. Furthermore, the disagreement on the most suitable sense of the target words between annotators is most complicated and unnoticed. While linguistics expert focus much more on syntactic structure and semantic content during annotation, laypersons lean on world knowledge in that context. This paper argues that meta language and world knowledge is a main influence to the annotation results, which should be taken into serious consideration during annotation. Thus, explicit crowd-sourcing and GWAP for Chinese WSD not only address solutions to quantity and efficiency problems, but also increases annotator diversification, native speaker instinct, thus might better reflect the nature feeling of Chinese native speakers.

References

[1] Kilgarriff, A., “English SENSEVAL resources in the public domain,” 1999. Available at: http://www.senseval.org/

Kilgarriff, A., “Lexicographical policy and procedure in the Hector project,” 1999. Available at: http://www.senseval.org/

[2] "中文詞彙網路 | Chinese Wordnet", lope.linguistics.ntu.edu.tw, [Online]. Available: http://lope.linguistics.ntu.edu.tw/cwn/query/. [Accessed: 21- Jul- 2016].

[3] Mihalcea, Rada, T.A. Chklovski, and A.Kilgarriff., "The Senseval-3 English lexical sample task,” Association for Computational Linguistics, 2004.

[4] "中研院平衡語料庫", asbc.iis.sinica.edu.tw, [Online]. Available: http://asbc.iis.sinica.edu.tw/. [Accessed: 21- Jul- 2016].

[5] "中文詞彙網路 | Chinese Wordnet", lope.linguistics.ntu.edu.tw, [Online]. Available: http://lope.linguistics.ntu.edu.tw/cwn/query/. [Accessed: 21- Jul- 2016].

[6] Doan, A., Ramakrishnan, R., & Halevy, A. Y., “Crowdsourcing systems on the world-wide web,” Communications of the ACM, 54(4), pp. 86-96, 2011.

[7] Bontcheva, K., Roberts, I., Derczynski, L., & Rout, D. P., “The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy,” In EACL, pp. 97-100, April 2014.

[8] Hong, J., & Baker, C. F., “How good is the crowd at real WSD?” In Proceedings of the 5th linguistic annotation workshop, pp. 30-37, Association for Computational Linguistics, June 2011.

[9] "Make your data useful | CrowdFlower", CrowdFlower, [Online]. Available: https://www.crowdflower.com/. [Accessed: 21- Jul- 2016].

[10] Von Ahn, L., & Dabbish, L., “Designing games with a purpose,” Communications of the ACM, 51(8), pp. 58-67, 2008.

[11] Chamberlain, J., Poesio, M., & Kruschwitz, U. “Phrase detectives: A web-based collaborative annotation game,” In Proceedings of the International Conference on Semantic Systems (I-Semantics’ 08), pp. 42-49, September 2008.

[12] Artignan, G., M. Hascoet, and M. Lafourcade, “Multiscale visual analysis of lexical networks,” In ¨13th International Conference on Information Visualisation, Barcelona, Spain, pp. 685–690, 2009.

[13] Venhuizen, N., Basile, V., Evang, K., & Bos, J., “Gamification for word sense labeling,” In Proc. 10th International Conference on Computational Semantics (IWCS-2013), pp. 397-403, 2013.

[14] Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., & Navigli, R., “Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose,” In ACL (1), pp. 1294-1304, 2014.

[15] Jurgens, D., & Navigli, R., “It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation,” Transactions of the Association for Computational Linguistics, 2, pp. 449-464, 2014.

[16] "Super Mario Bros. X - Home", supermariobrosx.org, [Online]. Available: http://www.supermariobrosx.org/. [Accessed: 21- Jul- 2016].

[17] Hsieh, Shu-Kai. Sense Structure in Cube: Lexical Semantic Representation in Chinese Wordnet. International Journal of Computer Processing Of Languages Vol. 23, No. 3. 243–253, 2011.