Word Sense Disambiguation via Human Computation

Nitin Seemakurty, Jonathan Chu, Luis von Ahn, Anthony Tomasic

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

{nseemaku,jechu}@andrew.cmu.edu {biglou,tomasic}@cs.cmu.edu

ABSTRACT

One formidable problem in language technology is the word sense disambiguation (WSD) problem: disambiguating the true sense of a word as it occurs in a sentence (e.g., recognizing whether the word "bank" refers to a river bank or to a financial institution). This paper explores a strategy for harnessing the linguistic abilities of human beings to develop datasets that can be used to train machine learning algorithms for WSD. To create such datasets, we introduce a new interactive system: a fun game designed to produce valuable output by engaging human players in what they perceive to be a cooperative task of guessing the same word as another player. Our system makes a valuable contribution by tackling the knowledge acquisition bottleneck in the WSD problem domain. Rather than using conventional and costly techniques of paying lexicographers to generate training data for machine learning algorithms, we delegate the work to people who are looking to be entertained.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge Acquisition H.5.3 [**HCI**]: Web-based interaction

1. INTRODUCTION

The human language is ambiguous. That is, words can be interpreted with different meaning depending on their surrounding context. Take, for example, the following two sentences [4]:

(a) I can hear *bass* sounds.

(b) They like grilled *bass*.

The word *bass* refers to a low-frequency tone in one sentence while it refers to a type of fish in the other. Although this sense recognition seems intuitive to humans, it is a much more sophisticated task for a machine, which has to cope with the unstructured nature of the data (language). This computational identification a word's meaning in a given context is called *Word Sense Disambiguation (WSD)*.

KDD-HCOMP'10, July 25, 2010, Washington, DC, USA. Copyright 2010 ACM 978-1-4503-0222-7 …$10.00

The relevance of WSD is becoming clear as advancing information/web technologies are catalysts for the production of enormous amounts of textual data, including articles, blogs, status messages, digitized books, etc. There is a growing need to introduce structure to this data in order to make it consumable and manageable by machines.

Current WSD algorithms use collections of data and machine learning algorithms to create models that determine the sense of the target word in the sentence. Generally supervised algorithms perform better than unsupervised algorithms [4]. These facts made human computation an ideal technique for this problem – with sufficient knowledge supervised algorithms can be used for almost all applications.

Currently knowledge acquisition for WSD is very expensive. Manual creation of a training dataset for a WSD system involves taking a large set of textual data, isolating words to disambiguate, and hand labeling each of these words with their gold label word sense. This process is an arduous and consequently an expensive one [3].

But what if we make this labeling process a pleasant one? This paper explores a new system: a game that is designed to capture human knowledge in a distributed fashion via an enjoyable game. Our study involves *assessing the effectiveness of this game in tackling the knowledge acquisition bottleneck*. Many elements of our game, named *Jinx,* are derived from a predecessor: the ESP Game [6].

1.1 Open Mind Initiative

Like the ESP Game, our game is much in tune with the efforts of the Open Mind Initiative [5], which focuses on collecting data from internet users in order to train machine learning algorithms. Our game is similar in that it attempts to use the efforts of regular internet users to tag the senses of words. However, as with the ESP Game, we place particular emphasis on the playability (i.e. viability) of our system.

2. GENERAL GAME PLAY

Jinx is an online cooperative two player game. When a player begins the game, he/she is anonymously paired with another random player. The anonymous pairing of players in most cases prevents any form of (cheating) communication between the two players. Each player interacts with the game independently. The players share only one aspect of the game: the current round. At any given time, both players view the same round, where a round is defined by a context (e.g. a sentence), and a highlighted word within that context.

The players are encouraged to rapidly type replacement words/phrases for the highlighted term. They are given incentive to type words that their partner is likely to type because both players are awarded points if and only if they both type the same string. As with the ESP Game, these players do not need to type their matching string at the exact same time, but both must have independently typed this string at some point during that round to receive points (see Figure 1 below).

We call the matching of two guesses for the same challenge a "tag". Once a tag is collected, the game awards points to each player and then proceeds to the next round. In the case where agreement cannot be reached, the round expires after 30 seconds. Players are presented rounds for exactly 3 minutes, and then they are taken to a summary page that recaps their performance and offers to restart the game with another anonymous player.

3. GAME DESIGN

Our game was originally designed to be more of a quiz comprised of a series of multiple choice questions. The player would be presented with a highlighted word in context, and then given multiple definitions to choose from. These definitions were intended to reflect different interpretations of the highlighted term. The player would be rewarded if their choice matched with the partner's choice. This setup, however, was inherently flawed.

(a) **Random guessing.** This design allowed players to collect points by blindly selecting answers and rapidly progressing through rounds, hoping for a lucky match with the partner. One solution could be to penalize for mismatch, but that would mean that a normal player would be deducted points due to the misbehavior of their partner.

(b) **Rigidity.** Limiting answer choices to a set of dictionary definitions also had the potential to create confusion if none of the definitions "worked". Players would eventually choose an answer that only weakly approximates the word's true meaning.

(c) **Playability.** This design requires a player to read through dictionary definitions. These definitions are surprisingly complex, subtle, and lengthy. Playing a single round required minutes of concentrated work that would quickly exhaust a player.

For these reasons, we adopted a more open-ended approach that allows players to rapidly guess word replacement strings. This approach minimizes a player's ability to match by randomly guessing. At the same time it makes the game more challenging and engaging by requiring that the players guess cooperatively, despite not being able to communicate. This cooperation emerges automatically from the task at hand and results in the generation of valuable tags.

3.1 Tag Quality

Because there are only a few possible word replacements that are reasonable given the sentence in any given round, players quickly recognize that making guesses from this limited set drastically increases their chances of matching with their partner. Consequently, the tags collected from the game are typically relevant word replacements.

Figure 1. Points are assigned to players only on a match. The number of points rewarded depends on several different factors.

3.2 INTERACTION DESIGN

In designing the point system for Jinx, we had several goals in mind [7]. We wanted to keep the game fast paced while still allowing for high quality input from players. Fast pace is encouraged because prior dry runs of the game indicate that the matching tag (i.e. the best replacement for the word) is commonly a very early guess during the round (one made quickly after reading the provided textual context). Giving players a sense of urgency encourages them to guess what is most intuitive to them, and this tends to be a successful tag. To generate this urgency, we reward each player $P = f(t)$ points upon a matching guess, where t is the number of seconds remaining in the round when the matching guess was made, and f is an increasing function (we currently use $f(t) = 10 * t$). The faster a player generates a tag, the more points the player is awarded. Notice that one of the players will inevitably make the matching guess before the other player does; each player is rewarded accordingly.

Players are also awarded bonus points for successfully matching with their partner on consecutive rounds. The value of the bonus increases on every round of their consecutive streak. At the end of a round, we award each player $g(n)$ bonus points, where n is the number of consecutive matches so far, and $g(n) = 10 * n$. Note that while g is a linear function, its effect is geometric because the total bonus a player receives is $10 * (1 + 2 + 3 + \cdots)$. This bonus system not only encourages players to keep playing, but also to keep playing with accuracy in mind. Most importantly, it keeps the game exciting.

4. GAME EVALUATION

To evaluate the game's correctness in collecting quality tags, we utilize data from the HECTOR project [1].

HECTOR provides us with a large set of contexts (usually complete sentences), each of which contains one demarcated word (we will hereafter refer to such contexts as *challenges*). Each challenge is also assigned a word sense, which corresponds to the particular sense that the challenge word carries in that particular sentence. Different challenges contain the same challenge word and yet can carry word sense.

To measure the game's ability to distinguish these alternate meanings, we injected a select subset of the HECTOR challenges into our backend and presented the game to a group of 11 players. The challenges were selected such that they involved only ten

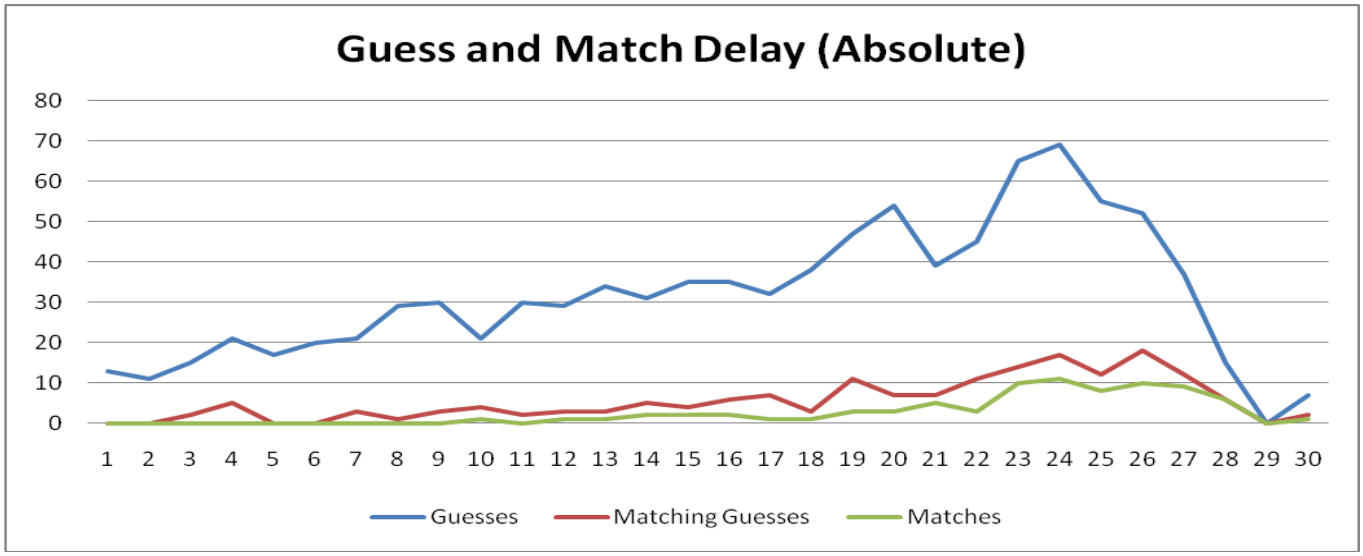

Figure 2: User Performance over Time

distinct challenge words but invoked multiple definitions of each of those words. The trial run lasted one hour.

4.1 Trial Run Observations

Several observations were recorded during the trial run:

1. The players' sentiment indicated that they found the game challenging, and consequently entertaining. While the players were keen to note several kinks in the implementation, most were intrigued by the concept.

2. Players felt that the game was too time constrained. The time limit for each round was 30 seconds. That is, players, once paired, had exactly 30 seconds to agree on a tag; otherwise, they were presented a new challenge. All participants in the study agreed that 30 seconds was too short a time period.

3. Interestingly, we were told by the players that, when presented an especially long challenge, they chose to read only the immediate context (5 to 10 words) surrounding the word. Much of the context provided by HECTOR is unnecessary to identify which meaning of a word is being invoked. Although presenting too little context has a chance of heightening the challenge word's ambiguity, it is clear that presenting less context than HECTOR offers can not only make the game less time constrained, but also make the game appear less formal and more fun.

Figure 2 is a graph of the activity of users over time. The "guesses" line (top) tracks the total number of guesses for a particular round. The "matching guess" line (middle) tracks the number of guesses that match *anyone* during the game. The "matches" line (bottom) tracks the number of guesses that match the paired player during the game. The data indicates a learning effect over time – players became more skilled at generating guesses and at generating guesses that matched. The drop off around round 25 was due to general fatigue and the trial session generally breaking up. This data also had one confounding factor – challenges were randomly drawn from a limited pool, so some players received the same challenge more than once.

4.2 Tag Analysis

In general the dataset produced by Jinx has interesting linguistic properties since it generates synonyms as perceived by the general public. Some guesses are wrong, but a match generally indicates a synonym of interest. For example, the synonym "bad" was generated for the word "bitter". This reasonable synonym is not listed in WordNet 3.0 [2].

To determine the usefulness of the dataset for WSD, for each challenge with a tag, we looked up the WordNet synonym sets ("synsets") for the word and then attempted to isolate a single synset using the tag. Each synset corresponds to a unique word sense. For each attempt, a tag is classified into one of five distinct categories. A tag may *uniquely* identify a synset word sense. A tag may correspond to *none* of the synsets. The word itself may be *missing* from WordNet (e.g., word *kneed*). The tag may *partially match* a synset (e.g. tag *orchestra* matches synset *dance orchestra*). Or, the tag may match multiple synsets and thus be *ambiguous*. Table 1 summarizes the results of this analysis.

Table 1: Summary of Tag Analysis

Result	*Count*	*Percent*	*Note*
Unique	45	54%	Unique to a word sense
None	24	29%	No word sense
Missing	2	2%	No word definition
Partial	3	4%	Partial match to sense
Ambiguous	9	11%	Two word senses
Total	83	100%	

The data indicates that 54% of the tags are uniquely associated with a word sense. This figure indicates that the game is relatively efficient in producing word sense labels for challenges. The three tags that partially appear in a synset also fall in this category. Some 29% of the tags did not directly correspond to a word sense. In subsequent versions of the game we plan to use *taboo words* [6] as a method of forcing players to attempt to think of other guesses. The same technique can be used to eliminate tags that ambiguously appear in more than one WordNet synset (9%). In the data, 2% of the words were not found in WordNet – for these words, additional word sense definitions must be constructed by hand using the tags – fortunately only a small percentage of words fall into this category.

A unique association does *not* mean that the associated word sense is *correct.* Of the 48 tags that are uniquely or partially associated with a word sense, fourteen had a single tag that covered all senses for a word, providing no discrimination power what-so-ever. Again, these tags are good candidates as taboo words. To evaluate the remaining words, we cross referenced the HECTOR gold label answer to the closest WordNet sense. Twenty-five tags (30%) mapped to the correct sense, nine tags (11%) mapped to the wrong sense.

5. CONCLUSION

In this paper we described a game, named *Jinx*, which is designed to generate word sense disambiguation (WSD) datasets. These datasets can be used to train high quality machine learning algorithms for the WSD problem.

The game accomplishes this task in a low cost, distributed, fashion by employing human beings to consider a word in sentence and generate guesses for synonyms for the word in the context of the sentence. The game uses a point system to provide utility to users and uses a cooperative, paired player, structure to make the game fun and to control the quality of the guesses.

We populated the game with ten words and multiple sentences from a widely recognized word sense evaluation dataset [1]. We then had people play the game for approximately an hour. In post-game interviews, most everyone reported that the game was fun. Thus we are confident that the game is capable of attracting a large, sustainable, audience.

Preliminary analysis of the guesses of the game indicates that many of the guesses correspond directly to synonym sets for words [2] in that context. Thus the game generates a set of synonyms to a particular word in a sentence *as perceived by the general public*. This dataset in itself contains interesting linguistic data. With respect to WSD, however, many guesses corresponded to more than one word sense or corresponded to incorrect word senses. We are currently exploring more sophisticated data analysis methods to extract a high quality WSD dataset.

6. ACKNOWLEDGMENTS

Thanks to Emily Leathers for work on this problem.

7. REFERENCES

[1] A. Kilgarriff. 1998. Senseval: An exercise in evaluating word sense disambiguation programs.

[2] Miller, G. A. 1995. WORDNET: A Lexical Database for English. Communications of ACM

[3] Ng, T. H. 1997. Getting serious about word sense disambiguation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington D.C.). 1–7.

[4] Roberto Navigli. 2009. Word sense disambiguation: a survey. ACM Computing Surveys, 41

[5] Stork, D. G. The Open Mind Initiative. IEEE Intelligent Systems & Their Applications, 14-3, 1999, pages 19-20.

[6] Von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In Proc. ACM CHI.

[7] Kraut, R. E. & Resnick, P. Evidence-based social design: Mining the social sciences to build online communities. Cambridge, MA: MIT Press. In preparation

[8] Zellweger, P.T., Bouvin, N.O., Jehøj, H., and Mackinlay, J.D. Fluid Annotations in an Open World. Proc. Hypertext 2001, ACM Press (2001), 9-18.