A Study of the Sense Annotation Process: Man v/s Machine.

Arindam Chatterjee, Salil Joshi, Pushpak Bhattacharyya
IIT Bombay
Powai,
Mumbai, 400076.
{arindam,salilj,pb}@cse.iitb.ac.in

Diptesh Kanojia
Gautam Budh Technical University
Lucknow,
Uttar Pradesh, 226021
dipteshkanojia@gmail.com

Akhlesh Kumar Meena
Akhil2068@gmail.com

Abstract

Does context help determine sense? This question might seem frivolous, even preposterous to anybody sensible. However, our long time research on Word Sense Disambiguation (WSD) shows that in almost all disambiguation algorithms, the sense distribution parameter *P(S/W)*, where *P* is the probability of the sense of a word *W* being *S*, plays the deciding role. The widely reported accuracy figure of around 60% for all-words-domain-independent WSD is contributed to mainly by *P(S/W)*, as one ablation test after another reveals.

The story with human annotation is different though. Our experience of working with human annotators who mark with WordNet sense ids, general and domain specific corpora brings to light the interesting fact that producing sense ids without looking at the context is a heavy cognitive load. Sense annotators do form hypothesis in their minds about the possible sense of a word ('most frequent sense' bias), but then look at the context for clues to accept or reject the hypothesis. Such clues are minimal, just one or two words, but are critical nonetheless. Without these clues the annotator is left in an indecisive state as to whether or not to put down the first sense coming to his mind. The task becomes all the more cognitively challenging, if the senses are fine grained and seem equally probable. These facts increase the annotation time by a factor of almost 1.5.

In the current paper we explore the dichotomy that might exist between machines and humans in the way they determine senses. We study the various parameters for WSD and also the sense marking behavior of human sense annotators. The observations, though not completely conclusive, establish the need for context for humans and that for accurate sense distribution parameters for machines.

1 Introduction

The process of sense annotation of words with senses is more accurate for humans than machines. The deciding parameter in the human sense disambiguation process is contextual evidence. Considering the principle of *weak AI*, the annotation procedure employed by the machine should make use of contextual evidence for disambiguation purposes in some form, which also conforms to the classical definition of WSD.

Our motivation is to exhibit that contextual evidence is a necessary attribute for the human tagging process. Without contextual information the human tagging process is crippled. Machines, which use the *P(S|W)* statistic for WSD, take human context-sensitive information to learn the *P(S|W)* measure. This is an adaptation of the contextual evidence used by human beings. Hence the principle of *weak AI* (Searle, 1980) holds for such WSD algorithms. Hence obtaining the *P(S|W)* values perfectly is of paramount concern for machines.

A glimpse at the history of the WSD task, reveals that the initial attempt was made towards WSD in the 1980s, when machine readable knowledge resources started becoming available, especially the Princeton WordNet (Fellbaum, 1998). In this period, context-based knowledge formed the sole tool for sense disambiguation purposes. In the 1990s statistical methods gained momentum, and till date have high accuracies in the sense disambiguation process (Ide and Véronis 1998).

Today, supervised approaches to WSD deliver far better results, compared to knowledge-based or unsupervised methods (Navigli, 2009). In a supervised framework, WSD is considered as a classification task, where senses of words are the classes. If we take a closer look at the state-of-the-art supervised algorithms for WSD, it will be evident that the parameters used by

such algorithms are mostly statistical, *i.e., corpus-based evidence*.

WSD researchers have tried to incorporate contextual support in the form of syntactical features, co-occurrence statistics and so on, but these algorithms do not perform significantly better over the Most Frequent Sense baseline.

A study of human cognition techniques in the annotation task unfolds that *context-based* evidence is a major parameter used by humans during annotation. In order to establish this, we made a study of the cognition techniques used by skilled lexicographers during the annotation task.

Consequently, state-of-the-art WSD algorithms use the *P(S|W)* statistic for annotation. In this paper we attempt to answer two basic questions regarding the annotation techniques of man and machine:

> ***For Humans**: Can humans annotate data efficiently without contextual evidence?*
> ***For Machines**: Do machines need context information during the annotation process?*

By providing relevant answers to these questions we intend to present a comparative study of methods employed by humans and machines for sense annotation.

The remainder of this paper is organized as follows. In section 2 we present related work. Section 3 presents the different corpora and annotation scenarios used in our experiments, followed by section 4 which discusses the supervised WSD algorithm which we use as a representation of the machine annotation strategy. Section 5 describes the need for a critical analysis of the algorithm and how the analysis is done. In section 6 we provide a layout of the experimental setup, followed by the results obtained and discussions in sections 7 and 8. Section 9 concludes the paper and section 10 pertains to future-work.

2 Related Work

In this paper we compare the annotation processes of state-of-the-art algorithms, which use the *P(S|W)* statistic as a classifier. (Lee, Ng and Chia 2004), (Khapra et. al 2010) are some examples of such state-of-the-art algorithms.

Unfortunately enough, our work seems to be a first of its kind, as to the best of our knowledge we do not know of any such work done before in the literature.

3 Corpora and Annotation Scenarios

Before elaborating on the details of the algorithm employed by us and the experiments conducted, it is essential to lay down an account of the types of corpora that have been used for our experiments and correspondingly the tagging techniques employed in each case.

It must be noted that from a linguistic point of view, the term *context* means *a set of surrounding words*. This can be a paragraph, sentence or a number of neighboring words depending on the need and focus of the experiment. In our case, we have considered the sentence surrounding the word as the context.

3.1 Context Sensitive Scenario

In this setting, a team of two skilled lexicographers was assigned the task of annotating the corpora from two *specific* domains (TOURISM and HEALTH) and a *generic* domain (NEWS), *using the context* of each word. This is a usual annotation scenario, where the lexicographer can *sense the context* and tag the word accordingly.

In order to enquire into the importance of context in the annotation processes of both human and machines, a scenario independent of contextual information was actuated.

3.2 Context Agnostic Scenario

In this setting, the same team of lexicographers was assigned the task of annotating the same corpora *without using the context*. To make the process more interesting and ensure genuineness, the corpora used in this case consisted of the list of unique words, obtained from the corpora used in the context *sensitive* scenario. The focus here was to make the lexicographers *agnostic* of the context.

3.3 Importance of Context in Annotation

After the annotation process, the lexicographers opined that this annotation task was cognitively taxing in the context agnostic scenario, which is a strong indication that context is the lone ingredient in the human annotation procedure.

In the case of machines, high accuracy WSD algorithms are mostly supervised and use the *P(S|W)* statistic for annotation. Besides, the *P(S|W)* statistic is obtained after training on a corpus in the context sensitive setting. Hence there is an absorption of contextual information in the generation of the *P(S|W)* values from the context sensitive training data.

4 WSD Algorithm: Iterative WSD

In order to compare the annotations of human and machine, the machine output WSD algorithm is necessary. For our experiments we have taken the output of a supervised WSD algorithm, developed at IIT Bombay, called *Iterative WSD (IWSD)* (Khapra et al. 2010). The algorithm is greedy and uses a scoring function to disambiguate senses. The scoring function, the parameters based on which IWSD has been designed and the basic algorithm are described in the following subsections.

4.1 Parameters for IWSD

Khapra et al. (2010) proposed a supervised algorithm for domain-specific WSD and showed that it beats the most frequent corpus sense and performs on par with other state-of-the-art algorithms like Personalized PageRank (Agirre, 2009). The various parameters used by Iterative WSD can be classified as:

Wordnet-dependent parameters

- *belongingness-to-dominant-concept*
- *conceptual-distance*
- *semantic-distance*

Corpus-dependent parameters

- *sense distributions*
- *corpus co-occurrences*.

4.2 Scoring function for IWSD

The scoring function of the IWSD algorithm integrates the WordNet-dependant parameters and the corpus-based parameters to rank the candidate senses of the target word. The scoring function is illustrated below:

$$S^* = \arg\max_i \theta_i V_i + \sum_{j \in J} W_{ij} V_i V_j \qquad (1)$$

Where,

J = *Set of disambiguated words*

θ_i = *BelongingnessToDominantConcept(Si)*

$Vi = P(S_i|word)$

$Wij = CorpusCooccurrence(S_i,S_j) * 1/WNConceptualDistance(S_i,S_j) * 1/WNSemanticGraphDistance(S_i,S_j)$

4.3 Algorithm

As stated earlier, IWSD is a greedy algorithm. The greedy nature of the algorithm can be explained through the steps followed by the algorithm.

Algorithm 1: *performIterativeWSD* (sentence)

1. Tag all monosemous words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
3. At each stage select that sense for a word which maximizes the score given by Equation 1

Monosemous words are used as the seed input for the algorithm but are not considered while calculating the precision and recall values. It is quite possible that a sentence may not contain any monosemous words in which case the algorithm will first disambiguate the least polysemous word in the sentence. In this case, the disambiguation will be performed only using the first term in the formula which represents the corpus bias (the second term will not be active as there are no previously disambiguated words).

The least polysemous word thus disambiguated will then act as the seed input to the algorithm. IWSD is clearly greedy. It bases its decisions on already disambiguated words, and ignores completely words with higher degree of polysemy. For example, while disambiguating bisemous words, the algorithm uses only the monosemous words and ignores completely the trisemous words and higher order polysemous words appearing in the context. This is illustrated in Figure 1.

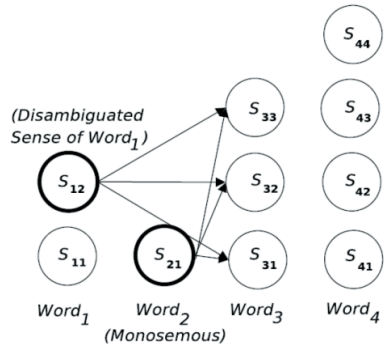

Figure 1*: IWSD Operation: Only previously disambiguated words and monosemous words are used while disambiguating Word3*

	Poly-semous words	Mono-semous words	Word-net Po-lysemy	Corpus Poly-semy
Noun	72225	61682	3.03	1.82
Verb	26436	4372	4.47	3.00
Adj	15462	30122	2.68	2.03
Adv	12907	10658	2.52	2.11
Overall	127030	106834	3.13	2.02

Table 1: Corpus statistics for NEWS domain

5 Critique of IWSD

The accuracy of the IWSD algorithm is comparable to other state-of-the-art supervised algorithms. Also IWSD has both statistical as well as contextual parameters in its scoring function. To get a deeper understanding of which parameters in the IWSD scoring function contribute towards its high accuracy, we performed the following tests.

5.1 Experiments on IWSD

First we conducted an ablation test on the parameters of the IWSD scoring function, tested on a generic corpus (NEWS), the details of which are given in table 1. The results of the ablation test are shown in table 2.

Ablation Parameter	Precision	Recall	F-Score
θ	79.61%	78.62%	79.11%
$P(S\|W)$	59.59%	58.84%	59.21%
Corpus-Cooccurence	79.57%	78.58%	79.07%
ConceptualDistance(S_i,S_j)	79.50%	78.51%	79.01%
SemanticSimilarity(S_i,S_j)	79.61%	78.62%	79.11%

Table 2: Results for ablation tests

We also compared the accuracy of IWSD against the *Most Frequent Sense (MFS) baseline,* on the NEWS corpus. MFS tags the words based on their $P(S_i|word)$ values. The results are shown in table 3. Next, in order to find the output of IWSD by assigning varying weights to the statistical and contextual parts of its scoring function, we tweaked the IWSD scoring formula into a linear combination of the statistical parameters and contextual parameters as explained in Equation 2, and tested for varying values of α. Table 4 shows the results of this experiment.

$$S^* = \alpha \arg\max_i \theta_i V_i + (1-\alpha)\sum_{j \in J} W_{ij} V_i V_j \quad (2)$$

sense in almost all cases, the accuracy will be bounded as follows:

	Precision	Recall	F-Score
MFS	79.57	78.52	79.04
IWSD	79.61	78.62	79.11

Table 3: MFS v/s IWSD

Alpha (α)	Precision	Recall	F-score
0	59.59%	58.84%	59.21
0.00001	79.48%	78.49%	78.98
0.0001	79.50%	78.51%	79
0.001	79.50%	78.51%	79.01
0.01	79.61%	78.62%	79.01
0.1	79.61%	78.62%	79.11
0.2	79.61%	78.62%	79.11
0.25	79.61%	78.62%	79.11
0.5	79.61%	78.62%	79.11
0.75	79.61%	78.62%	79.11
1	79.59%	78.60%	79.1

Table 4: IWSD results over range of alpha values

The results of tables 3 strongly indicate that IWSD algorithm is marginally better than MFS. From tables 2 and 4 it is evident that $P(S|W)$ is the prime parameter for IWSD.

The fact to be noted here is that, even though the $P(S|W)$ statistic apparently seems context agnostic, but it cannot be ignored that this statistic is in fact learned from corpus which was annotated in a context sensitive fashion. Hence in a way, the $P(S|W)$ parameter in IWSD is an adaption of Human Context Sensitive annotations.

5.2 Accuracy estimation of IWSD

Consider a sample word W which appears N times in the corpus. If W has k senses, S_1^w, S_2^w, S_3^w, ..., S_k^w which occur in the corpus with probabilities, P_1^w, P_2^w, P_3^w, ..., P_k^w, respectively.

As W occurs N times in the corpus, the total no. of occurrences of S_i^w in the corpus, can be captured in the following formulation:

$$\#S_i^w = P_i^w * N$$

For an algorithm like IWSD, which tags all the occurrences of a word W with the most frequent

$\Rightarrow min_i \{ P_i^w \} \leq accuracy \leq max_i \{ P_i^w \}$

$\Rightarrow (1- min_i \{ P_i^w \}) \geq (1 - accuracy) \geq (1- max_i \{ P_i^w \})$

$\Rightarrow (1 - min_i \{P_i^w\}) \geq error \geq (1 - max_i \{P_i^w\})$

Now, let S_d^w be the most frequent sense for W with probability P_d^w.
$\Rightarrow P_d^w > P_i^w, \forall i \neq d$
$\Rightarrow max_i \{P_i^w\} = P_d^w$.
$\Rightarrow \% error \geq (1 - P_d^w) * 100$

Since we have N occurrences of the word W in the corpus, the number of occurrences which will get tagged incorrectly will be at least,

$$N * (1 - P_d^w) \quad (3)$$

6 Experimental Setup

We report annotation experiments which were run on two specific domains (TOURISM and HEALTH) and a generic domain (NEWS). The TOURISM and HEALTH corpora consisted of around 8,000 words each and the NEWS corpus consisted of around 7,000 words.

To compare the annotated data obtained through the techniques described in Fig 1, we used *Jaccard's similarity coefficient and Cohen's Kappa coefficient.* We conducted the following experiments.

- We compared IWSD, Human Context Agnostic (HCA) and Human Context Sensitive (HCS) annotations taking HCS as the gold standard.
- We similarly compared the annotation genres mentioned above from the POS and ontological perspectives.

As described in section 3, we conducted experiments in both *context sensitive* and *context agnostic* scenarios to compare the annotation processes of man and machine. Sections 4 and 5 establish the algorithmic foundations of the IWSD. We can now categorize human and machine tagging into the genres illustrated in Fig 2.

Type of Experiment	Domain	POS Category				
		NOUN	ADJ	ADV	VERB	OVERALL
IWSD v/s HCA	TOURISM	0.34	0.13	0.05	0.31	0.27
	HEALTH	0.26	0.16	0.30	0.29	0.24
	NEWS	0.25	0.04	0.24	0.19	0.17

Table 5: Cohen's Kappa statistics for IWSD v/s Human Context Agnostic tagging

Human Context Agnostic (HCA)	Human Context Sensitive (HCS)
Machine Context Agnostic (MCA)	Machine Context Sensitive (MCS)

Figure 2: Human and Machine Tagging genres

7 Observations

7.1 Part-of-Speech (POS)-based and overall similarity measure

The similarity measures for Tourism, Health and News domains were calculated using Jaccard's similarity coefficient for all POS categories for every pair of annotation process as well as for IWSD. The Cohen's Kappa statistic was also calculated between Human Context Agnostic tagging and IWSD results. The results are summarized in tables 5 and 10.

7.2 Ontology-based similarity measure

Using Jaccard's similarity coefficient, the similarity measures for Tourism, Health and News domains were calculated for ontological categories for every pair of annotation process as well as for IWSD. We report the statistics for top few ontological categories that occur highest number of times. The results are summarized in tables 7 to 9.

8 Discussion

From the observations in the previous section, we see that for both the specific as well as generic domains, the similarity coefficients between pairs of annotations processes follows similar behavior. The similarity between IWSD and Human Context Sensitive tagging is highest among all three annotation comparisons. The lowest similarity occurs in case of Human Context Agnostic and Human Context Sensitive annotation pair. This behavior across all domains can be visualized as follows:

We observe similar behavior in the POS based as well Ontology based similarity measures, and the results are summarized in tables 5 to 10. The similarity measure is calculated using Human Context Sensitive data as gold standard.

We observed that the similarity between Human Context Agnostic annotations and Human Context Sensitive annotations is low across all domains, POS categories and ontological categories, which clearly indicate that the accuracy of human annotations get degraded significantly when humans try to annotate data without having knowledge of the context. Further, we also observed that extent of similarity between Human Context Agnostic annotations and Human Context Sensitive annotations was around 50%-60% across all domains, which is close to Wordnet First Sense baseline reported for these domains by Khapra et. al (2010).

Type of Experiment	TOURISM					HEALTH					NEWS				
	NOUN	ADJ	ADV	VERB	OVERALL	NOUN	ADJ	ADV	VERB	OVERALL	NOUN	ADJ	ADV	VERB	OVERALL
IWSD v/s HCA	0.72	0.56	0.61	0.74	0.68	0.69	0.56	0.79	0.69	0.67	0.66	0.40	0.75	0.53	0.62
HCS v/s IWSD	0.80	0.71	0.82	0.78	0.78	0.81	0.80	0.88	0.60	0.81	0.84	0.77	0.86	0.70	0.80
HCS v/s HCA	0.65	0.48	0.63	0.69	0.61	0.64	0.45	0.76	0.73	0.61	0.57	0.27	0.74	0.26	0.50

Table 6: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for HEALTH domain

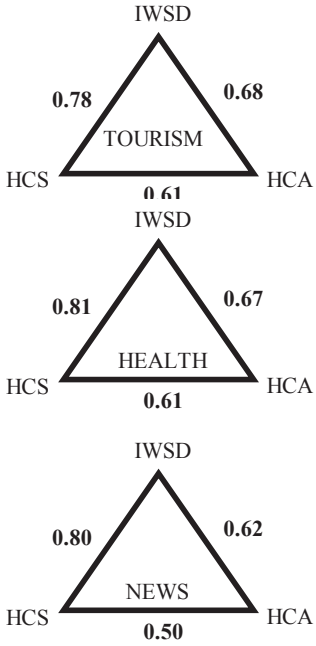

Figure 3: Visualization of comparison between tagging genres

In section 5, we have established that the prime parameter for IWSD is the *P(S|W)* statistic. Furthermore, we have also shown that IWSD is not truly context agnostic. The similarity measure is highest for IWSD and Human Context Sensitive annotations, which is consistent with our claim.

It can also be observed that similarity between Human Context Agnostic annotations and IWSD is low compared to similarity between Human Context Agnostic annotations and Human Context Sensitive annotations, which indicates that human beings are crippled without the context based knowledge during annotation. Conversely for machines, once training is done and the *P(S|W)* statistic is obtained, there is no further need of contextual evidence during annotation.

Ontological Category	TOURISM			
	Count	IWSD v/s HCA	HCS v/s IWSD	HCS v/s HCA
Verb of State	972	0.43	0.95	0.33
Action	863	0.25	0.83	0.21
Anatomical	798	0.35	0.89	0.34
Relational	721	0.33	0.75	0.18

Table 7: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for TOURISM domain

Ontological Category	NEWS			
	Count	IWSD v/s HCA	HCS v/s IWSD	HCS v/s HCA
Physical Place	2209	0.67	0.92	0.73
Person	1829	0.47	0.90	0.70
Artifact	1796	0.27	0.85	0.61
Bodily action	1582	0.20	0.83	0.55

Table 8: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for HEALTH domain

Ontological Category	HEALTH			
	Count	IWSD v/s HCA	HCS v/s IWSD	HCS v/s HCA
Bodily action	1198	0.06	0.95	0.89
Quantity	1188	0.01	0.90	0.16
Qualitative	1118	0.22	0.86	0.17
Numeral	1000	0.42	0.99	0.43

Table 9: Jaccard Similarity coefficient for highest occurring ontological categories across all pairs of annotation processes for NEWS domain

9 Conclusion

Based on our study on the two annotation scenarios in the two specific domains (TOURISM and HEALTH) and generic domain (NEWS), we conclude the following:

a) Contextual information is paramount for humans while disambiguating sense of a word.
b) The annotation process of tagging without the context is cognitively strenuous and time consuming as compared to tagging with help of the context.

c) In the case of machines, the *P(S|W)* measure can fetch high accuracies, provided that it has been correctly captured in the corpus by human beings, during annotation process. This in turn necessitates annotations with the help of context.
d) WSD algorithms, if trained on corpus generated through Context Agnostic annotation process, would result in low accuracies, as the *P(S|W)* parameter is not efficiently captured in this case.
e) Once the training process is over and *P(S|W)* statistic is captured, machines do not require further contextual information while annotating, unlike human annotation process. From this perspective, machines do not ape the human annotation technique but, through an adaptation of this technique provide high accuracies. Hence, machines conform to the principle of *weak AI* with respect to the annotation process.

10 Future-Work

In case of machines, we have observed that the *P(S|W)* statistic is the machine's adaption of human context sensitive annotation process and the principle of weak AI is satisfied here. However, the accuracies for WSD algorithms are not yet at par with human annotation quality. For this, we would like to see if using better contextual parameters in the IWSD scoring function and ranking the senses using a balanced formulation between statistical and contextual parameters, can further enhance the accuracy of the machine's annotation process.

In case of humans, a deeper insight into the exact cognitive processes which are involved during the annotation process could further leverage the study between man v/s machine sense annotation processes.

References

Eneko Agirre, and Soroa Aitor. 2009. *Personalizing pagerank for word sense disambiguation*. In EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 33–41, Morristown, NJ, USA. Association for Computational Linguistics.

Christian Fellbaum, 1998. *WordNet: An Electronic Lexical Database*

Nancy Ide and Jean Véronis. *Word Sense Disambiguation: The State of the Art Computational Linguistics*, 1998, 24(1). 2

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. *Domain-specific word sense disambiguation combining corpus based and wordnet based parameters*. In 5th International Conference on Global Wordnet (GWC2010).

Roberto Navigli. *Word sense disambiguation: A survey*. ACM Comput. Surv. 41, 2 (2009).

Lee, K. Yoong, Hwee T. Ng, and Tee K. Chia. 2004. *Supervised word sense disambiguation with support vector machines and multiple knowledge sources*. In Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 137–140.

Searle John. 1980. *Minds, brains, and programs*, Journal of Behavioral and Brain Sciences, Vol. 3, pages 417-424.