A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation

Hwee Tou Ng
Chung Yong Lim
Shou King Foo
DSO National Laboratories
20 Science Park Drive, Singapore 118230
{nhweetou,lchungyo,fshoukin}@dso.org.sg

Abstract

There is a general concern within the field of word sense disambiguation about the inter-annotator agreement between human annotators. In this paper, we examine this issue by comparing the agreement rate on a large corpus of more than 30,000 sense-tagged instances. This corpus is the intersection of the WordNet Semcor corpus and the DSO corpus, which has been independently tagged by two separate groups of human annotators. The contribution of this paper is two-fold. First, it presents a greedy search algorithm that can automatically derive coarser sense classes based on the sense tags assigned by two human annotators. The resulting derived coarse sense classes achieve a higher agreement rate but we still maintain as many of the original sense classes as possible. Second, the coarse sense grouping derived by the algorithm, upon verification by human, can potentially serve as a better sense inventory for evaluating automated word sense disambiguation algorithms. Moreover, we examined the derived coarse sense classes and found some interesting groupings of word senses that correspond to human intuitive judgment of sense granularity.

1 Introduction

It is widely acknowledged that word sense disambiguation (WSD) is a central problem in natural language processing. In order for computers to be able to understand and process natural language beyond simple keyword matching, the problem of disambiguating word sense, or discerning the meaning of a word in context, must be effectively dealt with. Advances in WSD will have significant impact on applications like information retrieval and machine translation.

For natural language subtasks like part-of-speech tagging or syntactic parsing, there are relatively well defined and agreed-upon criteria of what it means to have the "correct" part of speech or syntactic structure assigned to a word or sentence. For instance, the Penn Treebank corpus (Marcus et al., 1993) provides a large repository of texts annotated with part-of-speech and syntactic structure information. Two independent human annotators can achieve a high rate of agreement on assigning part-of-speech tags to words in a given sentence.

Unfortunately, this is not the case for word sense assignment. Firstly, it is rarely the case that any two dictionaries will have the same set of sense definitions for a given word. Different dictionaries tend to carve up the "semantic space" in a different way, so to speak. Secondly, the list of senses for a word in a typical dictionary tend to be rather refined and comprehensive. This is especially so for the commonly used words which have a large number of senses. The sense distinction between the different senses for a commonly used word in a dictionary like WordNet (Miller, 1990) tend to be rather fine. Hence, two human annotators may genuinely disagree in their sense assignment to a word in context.

The agreement rate between human annotators on word sense assignment is an important concern for the evaluation of WSD algorithms. One would prefer to define a disambiguation task for which there is reasonably high agreement between human annotators. The agreement rate between human annotators will then form the upper ceiling against which to compare the performance of WSD algorithms. For instance, the SENSEVAL exercise has performed a detailed study to find out the inter-annotator agreement among its lexicographers tagging the word senses (Kilgarriff, 1998c; Kilgarriff, 1998a; Kilgarriff, 1998b).

2 A Case Study

In this paper, we examine the issue of inter-annotator agreement by comparing the agreement rate of human annotators on a large sense-tagged corpus of more than 30,000 instances of the most frequently occurring nouns and verbs of English. This corpus is the intersection of the WordNet Semcor corpus (Miller et al., 1993) and the DSO corpus (Ng and Lee, 1996; Ng, 1997), which has been independently tagged with the refined senses of WordNet by two separate groups of human annotators.

The Semcor corpus is a subset of the Brown corpus tagged with WordNet senses, and consists of more than 670,000 words from 352 text files. Sense tagging was done on the content words (nouns, verbs, adjectives and adverbs) in this subset.

The DSO corpus consists of sentences drawn from the Brown corpus and the Wall Street Journal. For each word w from a list of 191 frequently occurring words of English (121 nouns and 70 verbs), sentences containing w (in singular or plural form, and in its various inflectional verb form) are selected and each word occurrence w is tagged with a sense from WordNet. There is a total of about 192,800 sentences in the DSO corpus in which one word occurrence has been sense-tagged in each sentence.

The intersection of the Semcor corpus and the DSO corpus thus consists of Brown corpus sentences in which a word occurrence w is sense-tagged in each sentence, where w is one of the 191 frequently occurring English nouns or verbs. Since this common portion has been sense-tagged by two independent groups of human annotators, it serves as our data set for investigating inter-annotator agreement in this paper.

3 Sentence Matching

To determine the extent of inter-annotator agreement, the first step is to match each sentence in Semcor to its corresponding counterpart in the DSO corpus. This step is complicated by the following factors:

1. Although the intersected portion of both corpora came from Brown corpus, they adopted different tokenization convention, and segmentation into sentences differed sometimes.
2. The latest version of Semcor makes use of the senses from WordNet 1.6, whereas the senses used in the DSO corpus were from WordNet 1.5.[1]

To match the sentences, we first converted the senses in the DSO corpus to those of WordNet 1.6. We ignored all sentences in the DSO corpus in which a word is tagged with sense 0 or -1. (A word is tagged with sense 0 or -1 if none of the given senses in WordNet applies.)

A sentence from Semcor is considered to match one from the DSO corpus if both sentences are exactly identical or if they differ only in the presence or absence of the characters "." (period) or "-" (hyphen).

For each remaining Semcor sentence, taking into account word ordering, if 75% or more of the words in the sentence match those in a DSO corpus sentence, then a potential match is recorded. These potential matches are then manually verified to ensure that they are true matches and to weed out any false matches.

Using this method of matching, a total of 13,188 sentence-pairs containing nouns and 17,127 sentence-pairs containing verbs are found to match from both corpora, yielding 30,315 sentences which form the intersected corpus used in our present study.

4 The Kappa Statistic

Suppose there are N sentences in our corpus where each sentence contains the word w. Assume that w has M senses. Let A be the number of sentences which are assigned identical sense by two human annotators. Then a simple measure to quantify the agreement rate between two human annotators is P_a, where $P_a = A/N$.

The drawback of this simple measure is that it does not take into account chance agreement between two annotators. The Kappa statistic κ (Cohen, 1960) is a better measure of inter-annotator agreement which takes into account the effect of chance agreement. It has been used recently within computational linguistics to measure inter-annotator agreement (Bruce and Wiebe, 1998; Carletta, 1996; Veronis, 1998).

Let C_j be the sum of the number of sentences which have been assigned sense j by annotator 1 and the number of sentences which have been assigned sense j by annotator 2. Then

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

where

$$P_e = \sum_{j=1}^{M} (\frac{C_j/2}{N})^2$$

and P_e measures the chance agreement between two annotators. A Kappa value of 0 indicates that the agreement is purely due to chance agreement, whereas a Kappa value of 1 indicates perfect agreement. A Kappa value of 0.8 and above is considered as indicating good agreement (Carletta, 1996).

Table 1 summarizes the inter-annotator agreement on the intersected corpus. The first (second) row denotes agreement on the nouns (verbs), while the last row denotes agreement on all words combined. The average κ reported in the table is a simple average of the individual κ value of each word.

The agreement rate on the 30,315 sentences as measured by P_a is 57%. This tallies with the figure reported in our earlier paper (Ng and Lee, 1996) where we performed a quick test on a subset of 5,317 sentences in the intersection of both the Semcor corpus and the DSO corpus.

[1] Actually, the WordNet senses used in the DSO corpus were from a slight variant of the official WordNet 1.5 release. This was brought to our attention after the public release of the DSO corpus.

Type	Num of words	A	N	P_a	Avg κ
Nouns	121	7,676	13,188	0.582	0.300
Verbs	70	9,520	17,127	0.556	0.347
All	191	17,196	30,315	0.567	0.317

Table 1: Raw inter-annotator agreement

5 Algorithm

Since the inter-annotator agreement on the intersected corpus is not high, we would like to find out how the agreement rate would be affected if different sense classes were in use.

In this section, we present a greedy search algorithm that can automatically derive coarser sense classes based on the sense tags assigned by two human annotators. The resulting derived coarse sense classes achieve a higher agreement rate but we still maintain as many of the original sense classes as possible. The algorithm is given in Figure 1.

The algorithm operates on a set of sentences where each sentence contains an occurrence of the word w which has been sense-tagged by two human annotators. At each iteration of the algorithm, it finds the pair of sense classes C_i and C_j such that merging these two sense classes results in the highest κ value for the resulting merged group of sense classes. It then proceeds to merge C_i and C_j. This process is repeated until the κ value reaches a satisfactory value κ_{min}, which we set as 0.8.

Note that this algorithm is also applicable to deriving any coarser set of classes from a refined set for any NLP tasks in which prior human agreement rate may not be high enough. Such NLP tasks could be discourse tagging, speech-act categorization, etc.

6 Results

For each word w from the list of 121 nouns and 70 verbs, we applied the greedy search algorithm to each set of sentences in the intersected corpus containing w. For a subset of 95 words (53 nouns and 42 verbs), the algorithm was able to derive a coarser set of 2 or more senses for each of these 95 words such that the resulting Kappa value reaches 0.8 or higher. For the other 96 words, in order for the Kappa value to reach 0.8 or higher, the algorithm collapses all senses of the word to a single (trivial) class. Table 2 and 3 summarizes the results for the set of 53 nouns and 42 verbs, respectively.

Table 2 indicates that before the collapse of sense classes, these 53 nouns have an average of 7.6 senses per noun. There is a total of 5,339 sentences in the intersected corpus containing these nouns, of which 3,387 sentences were assigned the same sense by the two groups of human annotators. The average Kappa statistic (computed as a simple average of the Kappa statistic of the individual nouns) is 0.463.

After the collapse of sense classes by the greedy search algorithm, the average number of senses per noun for these 53 nouns drops to 4.0. However, the number of sentences which have been assigned the same coarse sense by the annotators increases to 5,033. That is, about 94.3% of the sentences have been assigned the same coarse sense, and that the average Kappa statistic has improved to 0.862, signifying high inter-annotator agreement on the derived coarse senses. Table 3 gives the analogous figures for the 42 verbs, again indicating that high agreement is achieved on the coarse sense classes derived for verbs.

7 Discussion

Our findings on inter-annotator agreement for word sense tagging indicate that for average language users, it is quite difficult to achieve high agreement when they are asked to assign refined sense tags (such as those found in WordNet) given only the scanty definition entries in the WordNet dictionary and a few or no example sentences for the usage of each word sense. This observation agrees with that obtained in a recent study done by (Veronis, 1998), where the agreement on sense-tagging by naive users was also not high. Thus it appears that an average language user is able to process language without needing to perform the task of disambiguating word sense to a very fine-grained resolution as formulated in a traditional dictionary.

In contrast, expert lexicographers tagged the word sense in the sentences used in the SENSEVAL exercise, where high inter-annotator agreement was reported. There are also fuller dictionary entries in the HECTOR dictionary used and more examples showing the usage of each word sense in HECTOR. These factors are likely to have contributed to the difference in inter-annotator agreement observed in the three studies conducted.

We also examined the coarse sense classes derived by the greedy search algorithm. We found some interesting groupings of coarse senses for nouns which we list in Table 4.

From Table 4, it is apparent that the greedy search algorithm can derive interesting groupings of word senses that correspond to human intuitive judgment of sense granularity. It is clear that some of the disagreement between the two groups of human annotators can be attributed solely to the overly refined senses of WordNet. As an example, there is a total

loop: let $C_1, \ldots, C_M$ denote the current M sense classes
$\kappa^* \leftarrow -\infty$
for all i, j such that $1 \leq i < j \leq M$
let $C'_1, \ldots, C'_{M-1}$ denote the resulting $M-1$ sense classes by merging C_i and C_j
compute $\kappa(C'_1, \ldots, C'_{M-1})$
if $\kappa(C'_1, \ldots, C'_{M-1}) > \kappa^*$ then
$\kappa^* \leftarrow \kappa(C'_1, \ldots, C'_{M-1})$; $i^* \leftarrow i$; $j^* \leftarrow j$
end for
merge the sense class C_{i^*} and C_{j^*}
$M \leftarrow M-1$
if $\kappa^* < \kappa_{min}$ goto **loop**

Figure 1: A greedy search algorithm

Type	Avg Num of senses	A	N	P_a	Avg κ
Before	7.6	3,387	5,339	0.634	0.463
After	4.0	5,033	5,339	0.943	0.862

Table 2: Inter-annotator agreement for 53 nouns before and after the collapse of senses

Sense 1 change, alteration, modification – (an event that occurs when something passes from one state or phase to another: "the change was intended to increase sales"; "this storm is certainly a change for the worse")
Sense 2 change – (a relational difference between states; esp between states before and after some event: "he attributed the change to their marriage")
Sense 3 change – (the act of changing something; "the change of government had no impact on the economy"; "his change on abortion cost him the election")
Sense 4 change – (the result of alteration or modification; "there were marked changes in the lining of the lungs"; "there had been no change in the mountains")
Sense 5 change – (the balance of money received when the amount you tender is greater than the amount due; "I paid with a twenty and pocketed the change")
Sense 6 change – (a thing that is different; "he inspected several changes before selecting one")
Sense 8 change – (coins of small denomination regarded collectively; "he had a pocketful of change")

Figure 2: Seven senses of the noun "change" used by the human annotators

of 111 sentences in the intersected corpus containing the noun with root word form "change". They are assigned one of the seven senses listed in Figure 2 by the two groups of human annotators.

Based on the initial word senses assigned, $P_a = 0.38$ and $\kappa = -0.09$. (κ is negative when there is systematic disagreement.) However, the greedy search algorithm collapses sense 1, 2, 3, 4 and 6 into one coarse sense and sense 5 and 8 into another coarse sense. As a result, $P_a = \kappa = 1$, indicating perfect agreement when the senses are collapsed in the manner found. This corresponds to our intuitive judgment of the relative closeness of the various senses here.

Similarly, some of the 96 words for which the greedy search algorithm collapses into one single sense are such that the various senses are too close to be reliably distinguished. In short, we believe that the coarse sense classes derived by the greedy search algorithm, upon verification by human, can potentially serve as a better sense inventory for evaluating automated word sense disambiguation algorithms.

8 Related Work

Recently, both Bruce and Wiebe (1998) and Veronis (1998) have looked into algorithms to automatically generate better sense classes in a corpus-based, data-driven manner. However, the algorithms they used differ from ours. Bruce and Wiebe (1998) made use of an EM algorithm via a latent class model to derive better sense classes. Veronis (1998) performed a Multiple Correspondence Analysis on the table of annotations (a triple composed of a context, a judge and a sense) to reduce dimensionality followed by tree-clustering. In contrast, our greedy search algorithm is a simple but effective method that makes use of the Kappa statistic to search the space of possible sense groupings directly.

9 Conclusion

In this paper, we examined the issue of inter-annotator agreement on word sense tagging and presented a greedy search algorithm capable of generating coarse sense classes based on the sense tags

Type	Avg Num of senses	A	N	P_a	Avg κ
Before	12.8	5,115	8,602	0.595	0.441
After	5.6	8,042	8,602	0.935	0.852

Table 3: Inter-annotator agreement for 42 verbs before and after the collapse of senses

Noun	Coarse senses
air	wind/gas vs aura/atmosphere
board	committee vs plank
body	physical/natural object vs group/collection
change	modification vs coins
country	nation vs region/countryside
course	class vs action vs direction
field	land vs subject
foot	human body part vs unit vs lower part/support
force	strength vs personnel
light	lumination vs perspective
matter	concern/issue vs substance
party	political party vs social gathering vs group

Table 4: Coarse senses derived by the greedy search algorithm

assigned by two human annotators. We found interesting groupings of word senses that correspond to human intuitive judgment of sense granularity.

References

Rebecca Bruce and Janyce Wiebe. 1998. Word-sense distinguishability and inter-coder agreement. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing.*

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Adam Kilgarriff. 1998a. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language.*

Adam Kilgarriff. 1998b. Inter-tagger agreement. In *Advanced Papers of the SENSEVAL Workshop.*

Adam Kilgarriff. 1998c. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of LREC.*

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303–308.

George A. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–47.

Hwee Tou Ng. 1997. Exemplar-based word sense disambiguation: Some recent improvements. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 208–213.

Jean Veronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Advanced Papers of the SENSEVAL Workshop.*