Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation

Timothy Chklovski
Information Sciences Institute
University of Southern California
timc@isi.edu

Rada Mihalcea
Department of Computer Science
University of North Texas
rada@cs.unt.edu

Abstract

It is generally agreed that the success of a Word Sense Disambiguation (WSD) system depends, in large, on having enough sense annotated data available at hand, and a well-motivated sense inventory into which the disambiguations are made.

We report a Web-based approach to (1) constructing large sense tagged corpora by exploiting agreement of Web users who contribute word sense annotation, and (2) deriving a coarse-grained sense inventory from a fine-grained inventory by exploiting disagreements of independent contributors about word senses. We investigate the quantity and quality of the sense tagged data collected with this approach over the past year. We also present and evaluate an automatic clustering algorithm able to derive sense clusters that compare well with manually constructed clusters.

1 Introduction

One notoriously difficult problem in understanding text is Word Sense Disambiguation (WSD). Ambiguity is very common, especially among the most frequent words. Humans, however, are so competent at figuring out word senses from context that they usually do not even notice the ambiguities. While a large number of efficient WSD algorithms have been designed and implemented to date within the recent SENSEVAL evaluation frameworks and elsewhere, recently there has not been much progress on two related problems which are known to have a strong impact on the quality of WSD systems.

One such problem is the availability of sense tagged data. With a handful of tagged texts currently available, existing WSD systems are able to deal only with few pre-selected words for which hand annotated data was provided. The amount of sense tagged data available to date is limited to annotated examples for at most 300–400 words,[1] out of the total of 20,000 words that carry more than one possible meaning (in English).

The second problem is being able to make consistent and reasonable sense distinctions. Fine grained sense inventories such as WordNet make it hard even for humans to reliably and consistently distinguish among word senses. For example, the description of the Senseval-2 lexical sample task (Kilgarriff 02) compared WordNet and HECTOR sense inventories and concluded: *An implication for future WSD research is that it is time to turn our attention from algorithms, and to sense distinctions.*

The rest of the paper is organized as follows. In Section 2, we describe the system used to collect annotated data from Web users. In Section 3 and 4 respectively, we investigate the quantity and quality of data collected over the Web, and evaluate the WSD performance that can be achieved by relying on these data. In Section 5 we present how the coarse-grained clustering is created and its evaluation. We summarize our contributions in Section 6.

2 Building Sense Tagged Corpora with the Help of Web Users

To overcome the current lack of sense tagged data and the limitations imposed by the creation of such data using trained lexicographers, we designed a system that enables the collection of semantically annotated corpora over the Web.

Sense tagged examples are collected using a Web-based application that allows contributors to annotate words with their meanings. Tagging is organized by word: for each ambiguous word for which a sense tagged corpus is desired, contributors are presented with a set of natural language (English) sentence-long contexts each of which includes an instance of the ambiguous word.

The overall process proceeds as follows. Initially, example sentences are extracted from a large textual corpus. If other training data is not available, a number of these sentences are presented to the users for tagging in *Stage 1*. Next, this tagged collection is used as training data, and active learning is used to identify in the remaining corpus the examples that are "hard to tag". These are the examples that are presented to the contributors for tagging in *Stage 2*.

For all tagging, users are asked to select the sense they find to be the most appropriate in a given sentence. The selection is made from a drop-down list containing all WordNet senses of the current word, plus two additional choices, "unclear" and "none of the above."

The results of any automatic classification or the

[1] See http://www.senseval.org for currently available sense tagged data.

classification submitted by other users are not presented so as to not bias the contributor's decisions. Based on early feedback from both researchers and contributors, a future version of the system will allow contributors to specify more than one sense for a given instance.

2.1 Source corpora

The corpus from which we currently draw instances for annotation is formed by mixing of three different corpora, namely the *Penn Treebank* corpus, the *Los Angeles Times* collection as provided during TREC conferences, and *Open Mind Common Sense*[2], a collection of about 500,000 commonsense assertions in English as contributed by volunteers over the Web (Singh 02). We are currently in the process of integrating the *British National Corpus*; we also plan to integrate the *American National Corpus* as soon as it becomes available.

2.2 Sense Inventory

The sense inventory used in the current system implementation is WordNet (Miller 95). Users are presented with the current sense definitions from WordNet, and asked to decide on the most appropriate sense in the given context. Future versions of the system may adopt a new sense inventory, or use the coarse sense classes derived from WordNet, since the current fine granularity of WordNet was occasionally a source of confusion for some contributors and sometimes discouraged them from returning to the tagging task.

3 Quantity and Quality of Web-based Sense Tagged Corpora

Collecting from the general public holds the promise of providing much data at low cost. It also raises the importance of two aspects of data collection: (1) ensuring contribution quality, and (2) making the contribution process engaging to the contributors.

To ensure contribution quality, we collect redundant tagging for each item. The system currently uses the following rules in presenting items to volunteer contributors:

Two tags per item. We keep presenting an item for tagging until two taggings for it have been obtained.

One tag per item per contributor. We allow contributors to submit tagging either anonymously or having logged in. Anonymous contributors are not shown any items already tagged by contributors (anonymous or not) from the same IP address. A logged in contributor is not shown items that this contributor has already tagged.

In less than one year since the beginning of the activity, we collected almost 100,000 individual sense tags from contributors. Of these, approximately 16,500 tags came from using the system in the classrooms as a teaching aid (the web site provides special features for this). Future rate of collection depends on the site being listed in various directories and on the contributor repeat visit rate. We are also experimenting with prizes to encourage participation.

We measured the quality of the collected data in two ways. One is *inter-tagger agreement* (including κ statistics), which measures agreement between the tags assigned to the same item by two different annotators. The other is *replicability*, which measures the degree to which an annotation experiment can be replicated. According to (Kilgarriff 99), the capability of recreating a set of annotated data is an even more telling indicator of annotation quality than inter-tagger agreement.

3.1 Inter-Tagger Agreement

We can directly compare the inter-tagger agreement obtained so far with the agreement figures previously reported in the literature. (Kilgarriff 02) mentions that for the SENSEVAL-2 nouns and adjectives there was a 66.5% agreement between the first two tags collected for each item. About 12% of that tagging consisted of multi-word expressions and proper nouns, which are usually not ambiguous, and which are not considered during our data collection process. So far we measured a 62.8% inter-tagger agreement for single word tagging, plus close-to-100% precision in tagging multi-word expressions and proper nouns (as mentioned earlier, this represents about 12% of the annotated data). This results in an overall agreement of about 67.3% which is reasonable and closely comparable with previous figures.

3.2 Kappa Statistic

In addition to raw inter-tagger agreement, the kappa statistic was also determined, which removes from the agreement rate the amount of agreement that is expected by chance (Carletta 96).

We measure two figures: *micro-average* κ, where number of senses, agreement by chance, and κ are determined as an average for all words in the set, and *macro-average* κ, where inter-tagger agreement, agreement by chance, and κ are individually determined for each of the 280 words in the set, and then combined in an overall average. With an average of five senses per word, the average value for the agreement by chance is measured at 0.20, resulting in a *micro-average* κ *statistic* of 0.58. For *macro-average* κ estimations, we assume that word senses follow the distribution observed in the Open Mind annotated data, and under this assumption, the *macro-average* κ is evaluated at 0.35.

Only few previous sense annotation experiments report on the κ statistic, and therefore it is hard to compare the values we obtain with previous evaluations.

[2] http://commonsense.media.mit.edu

It is generally assumed that agreement above 0.80 represents *perfect agreement*, 0.60-0.80 represents *significant agreement*, 0.40-0.60 represents *moderate agreement*, and 0.20-0.40 is *fair agreement*. While most NLP applications seek data with an agreement that is at least *significant*, this is rarely the case in the task of sense annotation. Previous semantic annotation experiments report a macro-average κ for nouns of 0.30 (Ng *et al.* 99), as measured on the intersection between SemCor and the DSO corpus; a value of 0.49 for the annotation of 36,000 word instances in a French corpus (Veronis 00); a value of 0.44 for the Spanish SENSEVAL-2 task (Rigau *et al.* 01) (for the last two κ values, it is not clear whether they were computed using *micro* or *macro* average)

We also measured the κ statistic on the corpus constructed for the 29 nouns in the English lexical sample task at SENSEVAL-2. The SENSEVAL-2 English lexical sample data was constructed following the principle of *tag until at least two agree*. To create a setting similar to the Open Mind collection process, in this evaluation we only consider the first two (chronologically) tags. With an agreement by chance determined based on sense distributions drawn from the corpus itself, the macro-κ statistic for this sense tagged corpus is measured at 0.62, and the micro-κ statistic is evaluated at 0.65. On the same noun set, the Open Mind data has a macro-κ value of 0.43, and a micro-κ of 0.55. While κ statistics for the SENSEVAL-2 data are clearly higher, the figures are not however directly comparable since (1) SENSEVAL-2 data also includes multi-word expressions, which are usually easy to identify, and lead to high agreement rates; and (2) in Open Mind, the instances to be tagged for this set of 29 nouns were selected using an *active learning* process, and therefore these instances are "hard to tag".

3.3 Replicability

To measure the replicability of the tagging process performed by Web users, we carried out a tagging experiment for which annotation performed by "trusted humans" already existed. We used the data set for the noun *interest*, made available by (Bruce & Wiebe 94). Because this 2,369-item data set was originally annotated with respect to LDOCE, we had to map the sense entries from LDOCE to WordNet in order to make a direct comparison with the data we collect. The mapping was straightforward with one exception: all six LDOCE entries mapped one-to-one onto WordNet senses. There was one additional WordNet entry not defined in LDOCE; for this entry we discarded all corresponding examples from the Open Mind annotation.

Next, we identified and eliminated all the examples in the corpus that contained collocations (e.g. *interest rate*); these examples accounted for more than 35% of the data. Finally, the remaining 1,438 examples were displayed on the Web-based interface for tagging.

Number of training examples	Precision		Error rate reduction
	baseline	STAFS	
any	63.32%	66.23%	9%
> 100	75.88%	80.32%	19%
> 200	63.48%	72.18%	24%
> 300	45.51%	69.15%	43%

Table 1: Precision and error rate reduction for various sizes of the training corpus.

Out of the 1,438 examples, 1,066 had two tags that agreed, therefore a 74% inter-annotator agreement for single words tagging.[3] Out of these 1,066 items, 967 had a tag that coincided with the tag assigned in the experiments reported in (Bruce & Wiebe 94), which leads to an 90.8% replicability for single words tagging (note that the 35% monosemous multi-word expressions are not taken into account by this figure). This is close to the 95% replicability scores mentioned in (Kilgarriff 99) for annotation experiments performed by lexicographers.

In all, robustness of our data is also corroborated by the experience of a similar volunteer contribution project (Singh 02), which observed that the rate of maliciously misleading or incorrect contributions has been surprisingly low.

4 Exploiting Agreement of Human Annotators for WSD

We also carried out two sets of WSD experiments to further evaluate annotation quality. For these experiments, we used the items for which two Web annotators agreed on the sense tag assigned. One set of experiments disambiguated a held out subset of the collected corpus, with evaluations performed using ten-fold cross validations runs. This is the *intra-corpus* experiment, were both training and test sets are from the same source. The second set of experiments involves *inter-corpora* evaluations, in which the training corpus provided for the SENSEVAL-2 evaluation exercise is augmented with the examples contributed by Web users, and the performance is subsequently tested on the SENSEVAL-2 test data.

4.1 Intra-corpus WSD

In this experiment, we employ STAFS, one of the best performing WSD systems at SENSEVAL (Mihalcea 02). In current experiments, we use only a small set of features, consisting of the target word itself, its part of speech, and a surrounding context of two words and their corresponding parts of speech. The WSD performance is evaluated during 10-fold cross validation runs. We also compute a simple baseline, consisting of a simple heuristic that assigns by default the most frequent sense (also computed during 10-fold cross val-

[3] Addition of the 35% monosemous multi-word expressions tagged with 100% precision leads to an overall 83% inter-tagger agreement for this particular word.

Word	Set size	Baseline	WSD	Word	Set size	Baseline	WSD	Word	Set size	Baseline	WSD
activity	103	90.00%	90.00%	arm	142	52.50%	80.62%	art	107	30.00%	63.53%
attitude	107	100.00%	100.00%	bank	160	91.88%	91.88%	bar	107	61.76%	70.59%
bed	142	98.12%	98.12%	blood	136	91.05%	91.05%	brother	101	95.45%	95.45%
building	114	87.33%	88.67%	captain	101	47.27%	48.18%	car	144	99.44%	99.44%
cell	126	89.44%	88.33%	chance	115	56.25%	81.88%	channel	103	84.62%	86.15%
chapter	137	68.50%	71.50%	child	105	55.33%	84.67%	circuit	197	31.92%	45.77%
coffee	115	95.00%	95.00%	day	192	34.76%	44.76%	degree	140	71.43%	82.14%
device	106	98.12%	98.12%	doctor	133	100.00%	100.00%	dog	130	100.00%	100.00%
door	112	54.62%	45.38%	eye	117	96.11%	96.11%	facility	205	81.60%	74.40%
father	160	96.88%	96.88%	function	105	24.67%	32.00%	god	110	71.82%	81.82%
grip	239	45.94%	61.88%	gun	143	94.71%	94.71%	hair	147	96.67%	96.67%
horse	138	100.00%	100.00%	image	120	36.67%	71.67%	individual	103	96.15%	96.15%
interest	1066	39.91%	71.08%	kid	106	83.75%	84.38%	law	106	38.12%	66.88%
letter	137	85.00%	81.00%	list	102	100.00%	100.00%	material	196	77.60%	76.40%
mother	119	99.00%	99.00%	mouth	151	74.38%	77.50%	name	136	98.42%	98.42%
object	183	96.19%	96.19%	office	209	62.76%	61.03%	officer	103	56.15%	55.38%
people	120	99.17%	99.17%	plant	126	98.89%	98.89%	pressure	106	72.50%	70.62%
product	216	80.74%	81.48%	report	101	66.36%	60.91%	rest	360	51.11%	67.22%
restraint	204	22.92%	46.25%	room	124	100.00%	100.00%	sea	205	90.80%	90.80%
season	102	92.50%	92.50%	song	116	92.35%	92.35%	structure	112	75.38%	72.31%
sun	101	63.64%	66.36%	term	125	71.18%	90.59%	treatment	108	67.78%	66.67%
tree	105	100.00%	100.00%	trial	109	87.37%	86.84%	type	135	92.78%	92.78%
unit	108	54.44%	46.67%	volume	103	63.85%	54.62%	water	103	53.85%	72.31%

Table 2: Words with more than 100 sense tagged examples: (1) set size, (2) precision attainable with the most frequent sense heuristic, (3) precision attainable with the WSD system.

idation runs). Table 2 lists: all words for which we collected sense tagged data with at least 100 annotated examples available; the number of items with full inter-annotator agreement; the most frequent sense baseline; the precision achieved with STAFS.

For the total of 280 words for which data was collected from Web users, the average number of examples per word is 87. The most frequent sense heuristic yields correct results in 63.32% overall. When disambiguation is performed using STAFS, the overall precision is 66.23%, which represents an error reduction of about 9% with respect to the most frequent sense heuristic.

Moreover, the average for the 72 words which have at least 100 training examples (the words listed in Table 2) is 75.88% for the most frequent sense heuristic, and 80.32% when using STAFS, resulting in an error reduction of 19%. When at least 200 examples are available per word, the most frequent sense heuristic is correct 63.48% of the time, and the WSD system is correct 72.18% of the time, which represents a 24% reduction in disambiguation error. See Table 1 for precision and error rate reduction for various sizes of the training corpus.

For the words for which more data was collected from Web users, the improvement over the most frequent sense baseline was larger. This agrees with prior work by other researchers (Ng 97), who noted that additional annotated data is likely to bring significant improvements in disambiguation quality.

4.2 Inter-corpora experiments

In these experiments, we enlarge the set of training examples provided within the Senseval evaluation exercise with the examples collected from Web users, and evaluate the impact on performance of these new training examples. Only examples pertaining to single words are used (that is, we eliminate the SENSEVAL-2 examples pertaining to collocations).

There is only a small error rate reduction of 2% for fine grained scoring.[4] A more significant error reduction of 5.7% was observed for coarse grained scoring, found to be significant at a 0.2 level of significance using McNemar test. Notice that the examples used in our Web-based system are drawn from a corpus completely different than the corpus used for SENSEVAL-2 examples, and therefore the sense distributions are usually different, and often do not match the test data sense distributions (as is the case when train and test data are drawn from the same source). Previous word sense disambiguation experiments performed across diverse corpora have shown that variations in genre and topic negatively affect performance (Martinez & Agirre 00). The low error reductions obtained in our own inter-corpora experiments confirm these results.

5 Exploiting disagreement of human annotators to derive coarse sense clusters

For some items for which tags were collected, human annotators agreed – resulting in sense annotated data, but for other items they did not. We are exploiting these disagreements in a clustering method that automatically derives a coarse grained sense inventory, which compares well with manually constructed clusters.

The agreements/disagreements of human annotators can be reflected in a confusion matrix. As an example, Figure 1 shows the WordNet 1.7 senses of the word "presence" which were available to human annotators during the sense tagging. Figure 2 shows the resulting confusion matrix for this word.

[4] Fine grained scoring is a performance evaluation using word senses as defined in WordNet. Coarse grained scoring is an evaluation that relies on similar senses being grouped in clusters (e.g. by lexicographers).

1) **presence** - (a kind of *being*) – the state of being present; current existence; "he tested for the presence of radon"
2) **presence**, front - (a kind of *proximity*) – the immediate proximity of someone or something; "she blushed in his presence"; "he sensed the presence of danger"; "he was well behaved in front of company"
3) **presence** - (a kind of *spirit*) – an invisible spiritual being felt to be nearby
4) **presence** - (a kind of *impression*) – the impression that something is present; "he felt the presence of an evil force"
5) bearing, **presence**, mien, comportment - (a kind of *manner*) – dignified manner or conduct
6) **presence** - (a kind of *attendance*) – the act of being present

Figure 1: Six senses of the noun "presence" in WordNet 1.7.

	1	2	3	4	5	6	unclear	other
1	12	16	2	4	3	12	5	3
2		11	3	2	2	16	4	1
3			4	2		3	1	1
4				7	2	4	3	
5					6	8	3	
6						19	4	2
unclear								
other								

Figure 2: A confusion matrix (C) summarizing the tagging of the noun "presence." The rows and columns for tags "unclear" and "other" (unlisted sense) are presented here, but are not used in deriving coarse-grained sense inventory.

5.1 Computing similarity from confusion matrices

To cluster a fine-grained sense inventory, we first compute pairwise similarities of senses from the original confusion matrix. This step yields a matrix of pairwise similarities. As the next step, we apply a clustering algorithm that takes a matrix of pairwise similarities as input.

There are different ways to derive a similarity matrix from a confusion matrix. For example, (Godbole 02) utilizes the L_1 measure to assess similarity of two entries. That is,

$$Sim_{L_1}(i,j) = \sum_k |C_{ik} - C_{jk}| \tag{1}$$

where C_{ij} is the confusion matrix entry for senses i and j. However, this approach performs poorly without additional normalization when one sense has significantly more data than another. Alternatively, one could compute from the entries in the confusion matrix the conditional probabilities $p(j|i)$ of an item's second annotator tagging it with sense j given that the first annotator tagged it with sense i. Then, for each sense i we have a vector $\mathbf{P}_i$ of conditional probabilities $p(k|i)$, where k runs through all possible senses. Over these vectors, another approach widely used in information retrieval community — cosine similarity — could be deployed. That is, similarity of senses i and j could be measured by computing the inner product between $\mathbf{P}_i$ and $\mathbf{P}_j$.

Rather than use the above, we compute the similarity of two senses i and j in a way that is independent of entries that involve values other than i or j. Such local similarity measure allows us to avoid double-counting of similarity in later clustering of the senses.

Specifically, we compute similarity of i and j as the number of times these senses confused with each other, divided by the total number of times both annotators chose a sense from the set $\{i,j\}$ (the total number of times there could have been a confusion). For $i \neq j$ this definition can be stated as:

$$Sim_{ij} = \frac{C_{ij}}{C_{ii} + C_{ij} + C_{jj}}, \tag{2}$$

and for $i = j$ it simplifies to:

$$Sim_{ii} = \frac{C_{ii}}{C_{ii}} = 1. \tag{3}$$

When no data are available ($C_{ii} = C_{ij} = C_{jj} = 0$, and $i \neq j$) we presume that there is no similarity ($Sim_{ij} \stackrel{\text{def}}{=} 0$). When $C_{ii} = 0$ we still posit perfect self-similarity ($Sim_{ii} \stackrel{\text{def}}{=} 1$). This similarity measure has the following properties:

$$\begin{aligned} Sim_{ij} &\in [0,1], \\ Sim_{ij} &= Sim_{ji} \ (symmetry), \\ Sim_{ii} &\geq Sim_{ij}, \\ Sim_{ii} &= 1. \end{aligned}$$

Note that similarity of senses i and j does not constrain similarity of senses i and k or k and j, and thus a distance measure defined based on this similarity would not necessarily obey the triangle inequality.

5.2 Clustering

The task is to take similarity measures and create a partitioning of senses into clusters, each cluster representing a coarse sense.

Agglomerative clustering (see, for example, (Jain & Dubes 88; Rasmussen 92)) is a clustering method which iteratively finds two most similar clusters and merges them. At initialization, each element is treated as a cluster of one element. Throughout the clustering process, the similarity (or distance) between pairs clusters needs to be computed.

We use *complete linkage* agglomerative clustering, in which similarity between two clusters A, B is taken to be the *minimum* of all the distances between elements of the two clusters, $Sim(A,B) = min_{a \in A, b \in B} Sim(a,b)$. This requires no recalculation as clustering goes on, and is known to result in more spheroid clusters (in our case, the hope is that the clusters will be centered around the coarse meanings).

In addition to the issue of measuring the distance between clusters when merging them, there is also the issue of selecting a stopping criterion. The process of agglomerative clustering, if run to completion, will merge less and less similar clusters, terminating with all elements eventually being merged into a single large cluster.

We have chosen the following simple stopping criterion: of all the non-zero pairwise similarity weights, excluding self-similarities, we select the median one; we stop merging clusters when the similarity between the best candidates for a merge drops below this median similarity.

Figure 5.2 presents the symmetric matrix of similarities of senses of *presence*, computed from the confusion matrix according to Eqs. (2) and (3). Only the similarities exceeding the median cutoff are shown.

	1	2	3	4	5	6
1	1	0.410		0.174		0.279
2	0.410	1	0.167			0.348
3		0.167	1	0.154		
4	0.174		0.154	1	0.682	
5					1	0.242
6	0.279	0.348			0.242	1

Figure 3: Similarity of senses for the noun *presence*, derived from the confusion matrix in Figure 2 according to Eq. 2.

The resultant agglomerative clustering for the noun "presence" exactly matches the clustering provided by human experts:

$$Cl(\text{presence}) = \{\{1, 2, 6\}, \{3, 4\}, \{5\}\}.$$

5.3 Clustering Results

We present results of comparing our confusion matrix based clustering to the "correct," human-created clustering provided in conjunction with the Senseval competition. To further evaluate the performance, we also present two baselines obtainable by clustering elements randomly (with different assumptions). We also present the κ (kappa) values of the amount of improvement over the performance obtainable with random clusterings.

In the information retrieval community, important yardsticks for measuring the quality of results are precision and recall. They can be combined into a single score (the F-measure) in a principled way as detailed in (Van Rijsbergen 79, pp. 129–134). When applied to measuring the quality of a clustering (see, for example (Strehl 02, p. 109, eq. 4.24)), the F-measure (which takes an additional parameter β indicating relative importance of precision and recall) can be computed for a given "correct" cluster by computing how well the best attempted cluster replicates (captures) this correct cluster. The F-measure for the entire clustering is then computed by taking the weighted average of the F-measure for one correct cluster across all the correct clusters, as follows:

$$F_\beta = \sum_{C \in \mathcal{C}} \frac{\|C\|}{N_{tot}} \max_{A \in \mathcal{A}} \frac{(\beta^2 + 1)\|A \cap C\|}{\beta^2 \|C\| + \|A\|} \qquad (4)$$

where $\mathcal{C}$ is the set of clusters in the correct (human expert created) clustering, $\mathcal{A}$ is the set of clusters in the attempted clustering (which is being evaluated) and $N_{tot} = \sum_{C \in \mathcal{C}} \|C\|$ is the total number of elements (senses) being clustered. As is frequently done, we use the F_1-measure (i.e., $\beta = 1$), which weights precision and recall equally.

We compared our automatically derived clusters against manual coarse-grained clustering provided by lexicographers for the SENSEVAL-2 competition for the *English all words* and *English lexical sample* tasks.

The average F_1-measure of agreement between the agglomerative clustering and the manual clustering is 0.787. To better gauge this performance, we compare it to two baselines.

One baseline we compute is the "exhaustive clusterings" baseline. For a word with N senses, it calculates the average F_1-measure across all possible clusterings of N senses. Each clustering is treated as equally likely and each one is compared to the provided "correct" clustering $\mathcal{C}$. This baseline has the advantage of not being dependent on our particular attempted clustering, and thus can be used as a common baseline for other approaches.

Comparing solely to the "exhaustive clusterings" baseline, however, leaves open the possibility that clustering based on the confusion matrices outperforms the baseline by simply "guessing clusters of reasonable sizes." To investigate whether this is the case, we constructed a "same cluster size" baseline, employing a methodology similar to that of (Tomuro 01).

In the "same cluster size" baseline, clusters of the same size as in the attempted (confusion-matrix based) clustering are used, but senses are assigned to these clusters at random. For each word, 5000 same-cluster-size clusterings are generated and the mean F_1-measure for them is reported.

The "exhaustive clustering" baseline yielded an average F-measure of 0.666 across all words. The "same cluster size" baseline, which incorporated information about the attempted cluster sized, yielded an F-measure of 0.694. The F-measure for confusion matrix based clustering (0.787) improved on either baseline.

We present the detailed per-word results of the confusion-matrix based clustering, the two baselines, and the improvement provided by the clustering over the baselines in Table 3.

6 Summary

We proposed approaches to two problems in enabling WSD: the construction of large sense tagged corpora and the creation of coarse-grained sense inventories. In both cases, our solutions rely on data collected through a Web-based system where users can contribute their knowledge of sense annotations.

We evaluated the quantity and quality of the data collected from Web users, and showed how these data can be used to improve WSD performance. We also investigated methods of clustering fine-grained word senses based on confusion matrices of inter-annotator agreement.

The former result points to a promise of large, inexpensive, and potentially well-focused (through active

Word	F of agglom clust	F of same size baseline	F of exhaustive baseline	Word	F of agglom clust	F of same size baseline	F of exhaustive baseline	Word	F of agglom clust	F of same size baseline	F of exhaustive baseline
area	0.611	0.665	0.648	art	0.733	0.733	0.707	attention	0.556	0.621	0.657
authority	0.762	0.686	0.632	bank	0.773	0.620	0.575	bum	0.850	0.713	0.715
cell	0.900	0.733	0.654	chair	0.667	0.723	0.707	channel	0.771	0.674	0.628
child	0.833	0.721	0.707	church	0.778	0.778	0.767	circuit	0.667	0.702	0.657
claim	0.739	0.716	0.657	concentration	0.738	0.604	0.628	day	0.623	0.627	0.587
degree	0.905	0.662	0.628	detention	0.667	0.667	0.833	door	0.611	0.709	0.657
education	0.889	0.740	0.666	effect	0.722	0.756	0.657	example	0.548	0.666	0.642
extent	1.000	1.000	0.833	facility	1.000	0.702	0.670	family	0.694	0.586	0.615
fatigue	0.750	0.779	0.717	feeling	0.722	0.644	0.657	function	0.889	0.742	0.666
grip	1.000	0.726	0.640	growth	0.595	0.561	0.610	holiday	1.000	1.000	0.833
home	0.537	0.597	0.587	hope	0.722	0.656	0.646	importance	1.000	1.000	0.833
interest	0.633	0.590	0.640	lady	0.800	0.800	0.780	level	0.691	0.641	0.628
material	1.000	0.678	0.679	matter	0.849	0.652	0.651	meeting	0.706	0.653	0.646
mind	0.771	0.604	0.628	mouth	0.925	0.690	0.614	name	0.833	0.817	0.666
nation	0.750	0.775	0.717	nature	0.800	0.679	0.679	post	0.717	0.629	0.607
presence	1.000	0.652	0.646	process	0.694	0.650	0.654	reason	0.905	0.659	0.642
report	0.776	0.612	0.630	rest	1.000	0.844	0.648	restraint	0.694	0.650	0.654
school	0.762	0.672	0.628	sense	0.722	0.743	0.666	series	0.918	0.635	0.622
source	0.724	0.607	0.610	spade	0.778	0.778	0.767	story	0.905	0.655	0.651
stress	0.867	0.728	0.679	surface	1.000	0.671	0.654	term	0.667	0.705	0.657
test	0.571	0.644	0.628	text	1.000	0.720	0.715	type	0.778	0.634	0.651
unit	0.629	0.605	0.628	use	0.810	0.618	0.622	water	0.880	0.708	0.679
work	0.914	0.611	0.620					*AVERAGE*	0.787	0.694	0.666

Table 3: Per-word clustering results: (1) F-measure of correctness of agglomerative clustering; (2) baseline A: F of random clustering with clusters of same size as in (1); (3) baseline B: mean F across all possible clusterings.

learning) sense inventories; the latter result supports the broader notion that disagreement of human contributors, as well as their agreement, can carry useful, extractable information.

References

(Bruce & Wiebe 94) R. Bruce and J. Wiebe. Word sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 139–146, LasCruces, NM, June 1994.

(Carletta 96) J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

(Godbole 02) S. Godbole. Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers. Technical report, IIT Bombay, 2002. Available online at http://citeseer.nj.nec.com/godbole02exploiting.html.

(Jain & Dubes 88) A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, N.J., 1988.

(Kilgarriff 99) A. Kilgarriff. 95% replicability for manual word sense tagging. In *Proceedings of European Association for Computational Linguistics*, pages 277–278, Bergen, Norway, June 1999.

(Kilgarriff 02) A. Kilgarriff. English lexical sample task description. In *Proceedings of Senseval-2 Workshop, Association of Computational Linguistics*, pages 17–20, Toulouse, France, 2002.

(Martinez & Agirre 00) D. Martinez and E. Agirre. One sense per collocation and genre/topic variations. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, 2000.

(Mihalcea 02) R. Mihalcea. Instance based learning with automatic feature selection applied to Word Sense Disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-ACL 2002)*, Taipei, Taiwan, August 2002.

(Miller 95) G. Miller. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41, 1995.

(Ng 97) H.T. Ng. Getting serious about word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 1–7, Washington, 1997.

(Ng *et al.* 99) H.T. Ng, C.Y. Lim, and S.K. Foo. A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99)*, pages 9–13, College Park, Maryland, 1999.

(Rasmussen 92) E. Rasmussen. Clustering algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval Data Structures and Algorithms*. Prentice Hall, N.J., 1992.

(Rigau *et al.* 01) G. Rigau, M. Taule, A. Fernandez, and J. Gonzalo. Framework and results for the Spanish SENSEVAL. In *Proceedings of SENSEVAL-2, Association for Computational Linguistics Workshop*, Toulouse, France, 2001.

(Singh 02) P. Singh. The Open Mind Common Sense project. *KurzweilAI.net*, January 2002. Available online from http://www.kurzweilai.net/.

(Strehl 02) Alexander Strehl. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Unpublished PhD thesis, The University of Texas at Austin, May 2002. Avaliable online at: http://strehl.com/download/strehl-phd.pdf.

(Tomuro 01) Noriko Tomuro. Tree-cut and a lexicon based on systematic polysemy. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*, 2001. Available online at http://condor.depaul.edu/~ntomuro/research/naacl-01.html.

(Van Rijsbergen 79) C.J. Van Rijsbergen. *Information Retrieval*. London: Butterworths, 1979. available on-line at http://www.dcs.gla.ac.uk/ Keith/Preface.html.

(Veronis 00) J. Veronis. Sense tagging: Don't look for the meaning but for the use. In *Computational Lexicography and Multimedia Dictionaries (COMLEX'2000)*, pages 1–9, Kato Achia, Greece, 2000.