Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 18 (2004) 209–222
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Lexical knowledge and human disagreement on a WSD task

G. Craig Murray *, Rebecca Green

College of Information Studies, University of Maryland, College Park, MD 20742, USA

Received 6 October 2003; received in revised form 20 April 2004; accepted 5 May 2004
Available online 7 June 2004

Abstract

This paper explores factors correlating with lack of inter-annotator agreement on a word sense disambiguation (WSD) task taken from SENSEVAL-2. Twenty-seven subjects were given a series of tasks requiring word sense judgments. Subjects were asked to judge the applicability of word senses to polysemous words used in context. Metrics of lexical ability were evaluated as predictors of agreement between judges. A strong interaction effect was found for lexical ability, in which *differences* between levels of lexical knowledge predict disagreement. *Individual* levels of lexical knowledge, however, were not *independently* predictive of disagreement. The finding runs counter to previous assumptions regarding expert agreement on WSD annotation tasks, which in turn impacts notions of a meaningful ''gold standard'' for systems evaluation.

1. Introduction

In computational linguistics, word sense disambiguation often relies on human-annotated corpora for both training and evaluation. Typically these corpora are collections of text in which some number of the polysemous words have been annotated with pointers to specific word senses from a machine readable dictionary. A common approach to building a word-sense-disambiguated corpus is to have human judges read through passages that contain polysemous words and manually assign dictionary senses to those words. Judgments take into account the context surrounding the word, but retain a measure of subjectivity: Resolution of the ambiguity, while based on the context of use, is influenced by individual interpretation and knowledge of the language.

* Corresponding author. Tel.: +1-301-654-6781; fax: +1-301-314-9145.
E-mail addresses: gcraigm@umd.edu (G.C. Murray), rgreen@umd.edu (R. Green).

0885-2308/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2004.05.001

The solution to this dilemma has been to use multiple human judges on each passage, followed by an analysis of the agreement between judges. (This agreement has been variously referred to in the literature as inter-annotator agreement, inter-tagger agreement, or inter-assessor agreement.) By analyzing inter-annotator agreement, researchers have attempted to establish an upper bound of *human* performance in word sense disambiguation tasks.

Automated systems for word sense disambiguation (WSD) have compared their results to the conjectured upper bound of human performance as a "gold standard" (Kilgarriff, 1998). But the extent to which individuals agree when assigning dictionary senses to words in a corpus has varied widely in word sense disambiguation studies. Ahlswede and Lorand (1993) for example found inter-annotator agreement as low as 35%, while Kilgarriff and Rosenzweig (2000) achieved 95% inter-annotator agreement. While these inter-annotator agreement figures are not directly comparable because they stem from the use of different word sense inventories (Collins English Dictionary vs. HECTOR) such disparity has been taken to suggest that high levels of agreement are a function of training and/or lexical ability. Disagreements between judges on word sense assignments can be addressed through deliberation and an adjudicated result included in an annotated corpus. The extent of disagreements has been used in the past as a measure of confidence in the judgments made and as an indication of the upper bound for performance. Kilgarriff and Rosenzweig attribute their high numbers to the use of "trained" lexicographers. The inverse conjecture is that naïve annotators will produce much lower rates of agreement. The assumption has been that high rates of agreement equate to *accuracy* of sense assignment in some meaningful way. However, our evidence suggests that this is a misleading interpretation of agreement.

This paper addresses two assumptions common to word sense disambiguation tasks: (1) that experts (e.g., lexicographers) will agree more often with each other on lexical judgments than will non-experts, and (2) that the level of agreement achieved by experts constitutes some measure of "correctness". The research reported here investigates the interaction between lexical ability and lexical judgments. Our findings support the first assumption but challenge the second: Experts do tend to agree with each other more than non-experts do, but it is not absolute expertise alone that predicts the level of agreement, and agreement alone cannot be considered to convey correctness. Rather, lexical ability interacts with other factors and *differences* in ability are shown to be a stronger predictor of agreement than *levels* of ability.

We believe the operating assumption of agreement as a stand-alone measure is intrinsically flawed. Intuitively, we would expect an increase in the language expertise of judges to facilitate higher agreement measures for a lexical judgment task. However, there are other contributors to agreement. Fellbaum et al. (1997, 1998) found that "naïve" annotators given task-specific training were more likely to agree with one another than with their "expert" trainers. If level of knowledge were the only contributor to agreement, such an observation would be unlikely. Some obvious contributions to agreement are context of training, task specificity and granularity of judgments. Aspects of the dictionary used and the corpus selected also play a role. We believe, however, that lexical knowledge and problem solving strategies contribute far more to agreement between *novice judges* than has previously been considered. In any case, too little is known about why individuals agree or disagree on word sense judgments. Since the level of human agreement on such tasks has been used to frame our understanding of what "perfect" system performance would be, it is important that we learn more about what human agreement on a word sense disambiguation task means.

Because the primary focus of past studies has been on *machine* performance, investigation into *human* performance has been minimal. Some psycholinguistic research has touched on the issue. In a card-sorting task, Jorgensen (1990) identified a tendency in subjects to shift their classification of polysemous words when dictionary examples were available, suggesting a difference between dictionary-governed interpretation and naïve interpretation. Similarly, Véronis (1998) identified a granularity effect for both confidence and subjective agreement when using a dictionary with subtle differences in definition senses. Centrality of word meanings also has a significant effect. Williams (1992) found asymmetric relationships between synonyms of polysemous terms – central terms for a concept have a stronger influence on recognition of peripheral concepts than vice versa. So clearly there are factors that result from the *context* of the task and the *experience* of the judges, not just the *content* of the text.

What do individual judges bring with them? As a foundation for research, one of our goals should be to investigate empirically what some of the influences on inter-annotator agreement are. Why do only some annotators agree on classification of terms, and not others? Without this information we cannot really answer the question Ahlswede and Lorand (1993) asked a decade ago, ''Which of the human informants should the computer agree with, if the humans cannot agree among themselves?''.

As a first step in investigating these questions, we looked at human subjects' lexical knowledge and their performance on a word sense disambiguation task. We evaluated the ability of standardized testing to predict inter-annotator disagreement. We then compared agreement between pairs of individuals to agreement between single individuals and a gold standard used by the WSD community. Our findings indicate that agreement serves to describe the relative knowledge of the judges far more than it describes the ''correctness'' of their judgments.

2. Limitation of scope

There are limitations in this study worth noting: (a) The task of annotating text passages with particular entries or glosses in a machine readable dictionary is artificial. There is no evidence to suggest that individuals make distinctions of meaning delineated in this way outside the framework of such a task. (b) Within the framework of such a task, the original construction of the dictionary chosen plays a large role (Hanks, 2003; Atkins and Levin, 1991). (c) The textual contexts in the corpus are sometimes insufficient to clarify the use of the term being annotated (Krishnamurthy and Nicholls, 2000). (d) Senses appear in corpora for which simple mappings to available dictionaries do not exist. Texts often include metaphoric coercion of meaning for which designation of one sense over another does not clearly convey the meaning. Even when assignments of multiple meanings are allowed, the binding of the two meanings *via the context* can yield a unique semantic interpretation not captured in dictionaries. In order to explore specific factors such as individual lexical knowledge, we limited our study to conditions that parallel an existing defined task, namely human WSD judgments within the SENSEVAL-2 framework. A deeper exploration of the role of particular MRD's is warranted. For example the method of development and details of entry in a dictionary can have significant impact. These questions fall outside the scope of the present study. So too, the interaction between real-world knowledge (e.g., of world history) and the content of the corpus warrants investigation. This is also beyond our present scope.

3. Methodology and design

To assess factors contributing to inter-annotator agreement we conducted a test in two parts. The first part was an assessment of verbal knowledge. The second part was a word sense disambiguation task. A Web interface was developed to facilitate data collection and to control the flow of the user experience.

To assess subjects' knowledge of the English language, each test began with a section of multiple-choice questions taken from the verbal portion of a GRE practice exam (ETS, 1997). The full GRE is designed to test an individual's aptitude for graduate level study. Questions in the GRE have been vetted for their ability to identify a test subject's placement within a population. We use the questions in the verbal sections operationally as a reflection of subjects' lexicographic expertise relative to the general population. Twenty-seven questions were presented in a multiple-choice format. Seven of the questions asked the subject to select the best pair of words from a list to fill in blanks left in test sentences. Nine of the questions presented the subject with a pair of words and asked them to select another pair from a list of paired words that expressed the same relationship. The last twelve questions asked the subject to choose the correct antonym of a given word. The questions taken together are an entire section of a verbal GRE test with the reading comprehension portion removed. No other selection criteria were applied to reduce the number of questions. The correct answer and point value assigned to each question were taken from the ETS scoring key, which gives heavier weight to more difficult questions. The resulting scores are taken to reflect the subjects' knowledge of semantic and lexical relationships between terms in the English language, relative to a normalized population.

The second portion of the test presented the user with a series of English passages taken from the SENSEVAL-2 training corpus. The SENSEVAL-2 corpus includes context passages for ambiguous words and annotations marking head words with WordNet senses. (See Edmonds and Cotton, 2001 for an overview of SENSEVAL-2. For an overview of WordNet see Fellbaum, 1998.) We chose to select words that would accurately represent an entire training corpus rather than focus only on words of special interest, or words known to be problematic. Thirteen task words were randomly selected – six nouns, four verbs and three adjectives. The distribution of words across parts of speech in our selection reflects the distribution of nouns, verbs, and adjectives within the WordNet lexical database (Miller et al., 1990). Forty-seven context passages for those words were randomly selected from the SENSEVAL corpus. Each word was presented in at least three contexts, and no word was presented in more than six. Subjects were shown one passage at a time, grouped by word, with the ambiguous task-word highlighted in red.

Glosses taken from WordNet synsets were listed below each context passage presented. After reading each passage, users were asked to select the WordNet sense(s) they felt best applied to the highlighted word *as used in that context.* WordNet senses were used to facilitate comparison with the annotations in the SENSEVAL corpus. The presented order of senses (glosses) was randomized for each presentation of each context. Thus even when viewing senses for words they had seen in a previous context, the subjects were still more likely to read the full list of senses. This was done to eliminate any task presentation effect, e.g., from subjects simply choosing the first sense on the list. As a result, each user would disambiguate all test tasks for a given word, one after another, but would have to visually rescan the whole list of available senses each time. Subjects were encouraged to choose a single sense but were allowed to select more than one sense if they

felt more than one applied. After selecting one or more senses, the subject was asked to rate each of their sense selections both for applicability and for confidence.

Two seven-point Likert scales were presented for each answer. Likert scales (Likert, 1932) have been used to collect subjective judgments in a variety of tasks. Rorvig (1988) and Ellis (1996) have both demonstrated the suitability of Likert scales for capturing opinions regarding applicability of text, and several studies have indicated seven-point scales as appropriate to such tasks (e.g. Cuadra and Katter, 1967; Rong et al., 1998). The scale for applicability in our study ranged from "Barely applies" to "Applies perfectly". The scale for confidence ranged from "Others will agree" to "Others will disagree" (see Fig. 1 for detail). Subjects were permitted to make changes to their selections. The number of changes made was recorded as well as elapsed time per context and overall elapsed time.

4. Results

4.1. Subjects

Twenty-seven student subjects participated as judges in the study, fourteen undergraduate students and thirteen graduate students. Ages of subjects ranged from 18 to 60. With one exception, all of the subjects stated that they were native speakers of English; fifteen subjects also spoke another language. There was no particular selection criterion applied for disciplines of study. The distribution of students' majors is assumed to be representative of the general population of students at a public institution of higher education.

Data for each subject were collected in a controlled setting, stored in separate logs and tabulated. The average time for completion of the test was approximately 45 min.[1] The average number of senses per context selected by human judges was 1.24. Scores for the GRE-task were calculated against keys provided by the test authors. Elapsed time and number of changes were recorded, as well as selections made during the WSD task.

Because subjects were allowed to make multiple selections, and because they were asked to express the degree to which each selection applied, a simple hit or miss scoring approach is not an appropriate measure of "correctness". Moreover, we are not concerned here with right or wrong answers, but with the degree of disagreement between pairs of judges. For this reason we use a vector based distance measure, in which the disagreement between a set of judgments can be expressed as an absolute value of the distance between coordinates in a multi-dimensional space.

We used a measure proposed by Hripcsak et al. (1995) as a metric of the distance between judgments. Their study compared the judgments made by physicians, internists, radiologists, lay people, and a natural language processor. Subjects were asked to review patients' medical charts and assess whether certain medical conditions were present. The method allows for assignment of multiple medical conditions and does not require that all subjects make judgments on exactly the

[1] The time on task for word sense disambiguation makes up the bulk of this time with a mode near 30 min. The task contexts for SENSEVAL-1 and SENSEVAL-2 annotators differ from the task described here and direct comparisons to their time on task would be hard to interpret.

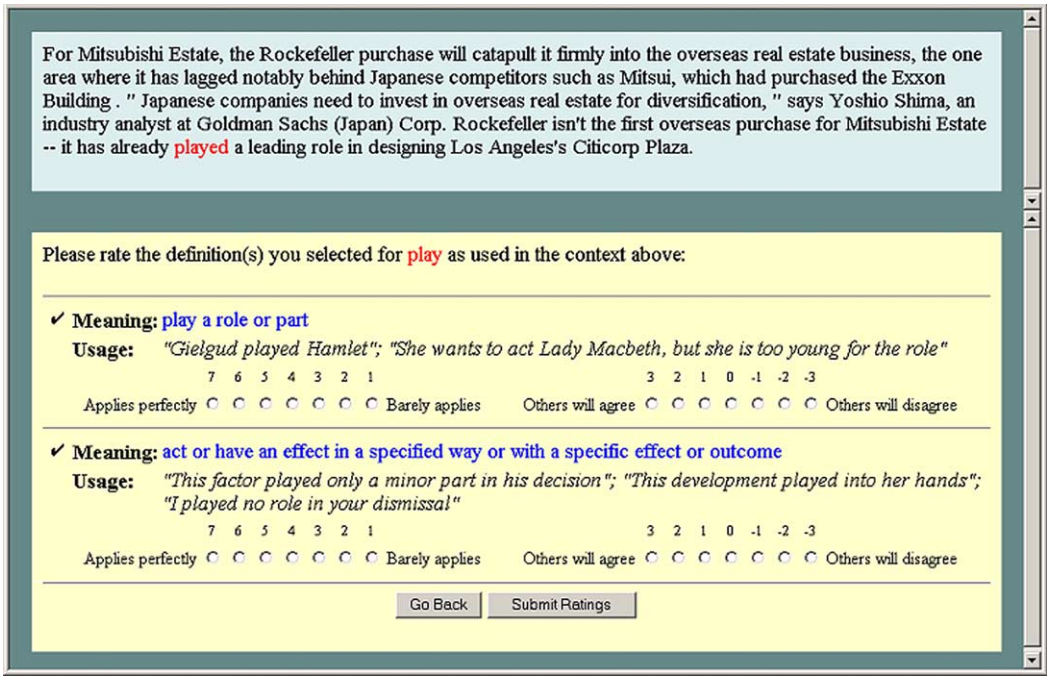

Fig. 1. Likert scales used to capture applicability judgments and confidence.

same number of charts. The one variation in our study is that rather than simply stating whether or not a condition applies (or in our case a word sense), we also have a judgment from the subject on *degree* to which it applies.

For each subject we have a vector of scores for each of the word-context tasks presented. The scores in each vector express the sense selections made by the subject. All senses presented to the subject are represented in the vector. Senses that were not selected by the subject are assigned a value of 0. Senses that were selected are given the value chosen by the subject from the 7-point Likert scale of applicability (as shown in Fig. 1). Scores for confidence were collected in the same way, but stored in separate vectors. Confidence scores were analyzed in the same manner as applicability scores, described below.

Borrowing from Hripcsak et al., we use the following notation to define the score values. Given a set of J subjects (denoted as $j = 1, 2, 3, \ldots, J$) and I tasks ($i = 1, 2, 3, \ldots, I$), a C-length vector X_{ij} contains values assigned to each task i by subject j, where C is the number of senses offered for that task's target word:

$$X_{ij} = \{x_{ij1}, x_{ij2} \ldots x_{ijC}\}.$$

Each score x_{ijc} is the degree to which subject j felt sense c applied to the target word in context i. Score values range from 0 to 7. We can then express the amount of disagreement between two judges j and k on context i as a distance d_{ijk} between vectors X_{ij} and X_{ik}

$$d_{ijk} = \sum_{c} |x_{ijc} - x_{ikc}|.$$

We calculate the absolute value of the differences of assertions made and treat this as a distance measure. The measure is weighted not only by the number of assertions made, but the degree of assertion. Pair-wise comparisons were made between each subject for each context task presented. The distance between subjects is then calculated as the average of distances across the number of tasks in common

$$\overline{d}_{jk} = \sum d_{ijk}/n_{jk}.$$

The reader is referred to Hripcsak et al. (1995) for the accompanying calculations of variance, covariance, and standard error.

In order to compare the subjects' scores to a SENSEVAL-2 "gold standard", we treated the SENSEVAL corpus as another subject in the study and calculated the same vectors and distance measures. The SENSEVAL answer tags within each context were extracted as the target senses. Scores for this gold standard were all given a value of 7 on the scale of applicability for assigned (target) senses and 0 for all other senses. Although probability weights are available for sense tags in the SENSEVAL corpus, they cannot be usefully applied in this context, and we chose not to incorporate them into our measure. The value of comparing our subjects' judgments to the SENSEVAL-2 "gold standard" lies in ranking the subjects' performance against one another, not in using the collective judgment expressed in the gold standard as a rubric. Our comparative measure focuses only on difference scores, and the subjects' relative rankings would not have been changed by incorporation of probability weights from the SENSEVAL corpus. Moreover, the relationship between probability and appropriateness is not direct. Measures of individual's statements of *applicability* as compared to the gold standard *probabilities* have questionable meaning. In the context of this study they lack a theoretical basis for interpretation.

Using the measures described, pair-wise comparisons were made between each pair of human judges and between each human judge and the SENSEVAL gold standard. Pair-wise comparisons between human subjects were made for both ratings of *applicability* of senses and ratings of *confidence*. Comparison between human subjects and the gold standard were made based on applicability scores only. For each pair of subjects we also calculated the difference between their lexical ability (GRE-task) scores. Correlation and regression analysis was performed on the scores calculated.

4.2. Correlations

Our analysis focuses on the factors contributing to pair-wise disagreement between subjects and to relative disagreement between subjects and the gold standard. Disagreement is calculated based on the vector distances described above. We also looked at correlations between types of scores for individual subjects, looking for interaction effects of GRE and WSD task scores. Level of education did not show a strong predictive effect and the GRE measure was taken to be a more explicit measure of applicable domain knowledge.

Of primary interest is the relationship between lexical knowledge scores and other parameters. Scores from the GRE task correlated negatively with measures of individual subjects' disagreement with the gold standard; where GRE-task scores were high, subjects were more

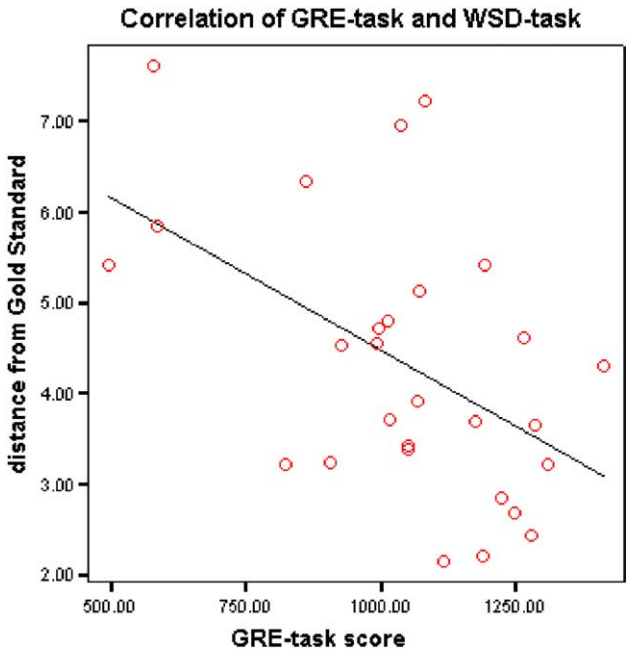

Fig. 2. Negative correlation between GRE task scores and disagreement with the gold standard.

likely to agree with senses in the gold standard (Fig. 2). This has serious implications for the selection of word sense annotators, which we discuss below. There was no correlation found between GRE-task scores and the amount of time individuals spent on the GRE task. Disagreement between individuals and the gold standard, however, did correlate negatively with *time* spent on the GRE task. That is, individuals who spent more time on the GRE task made judgments in the WSD task that were more like those in the SENSEVAL-2 corpus. This suggests an influence from mental processing separate from the influence of prior knowledge. Implications of these two findings are further discussed below. Table 1 shows correlations calculated using Pearson's r for distance from the gold standard (GS distance), GRE task scores, and time on task for the GRE.

A regression model and analysis of variance were used to analyze the interaction of factors contributing to pair-wise subject disagreement. The GRE-task scores for each subject in any pair

Table 1
Correlation matrix

	GRE-task	GRE-time
GS distance	−0.615**	−0.405*
GRE-task		0.100
$n = 26$		

* $p < 0.05$ (2-tailed).
** $p < 0.01$(2-tailed).

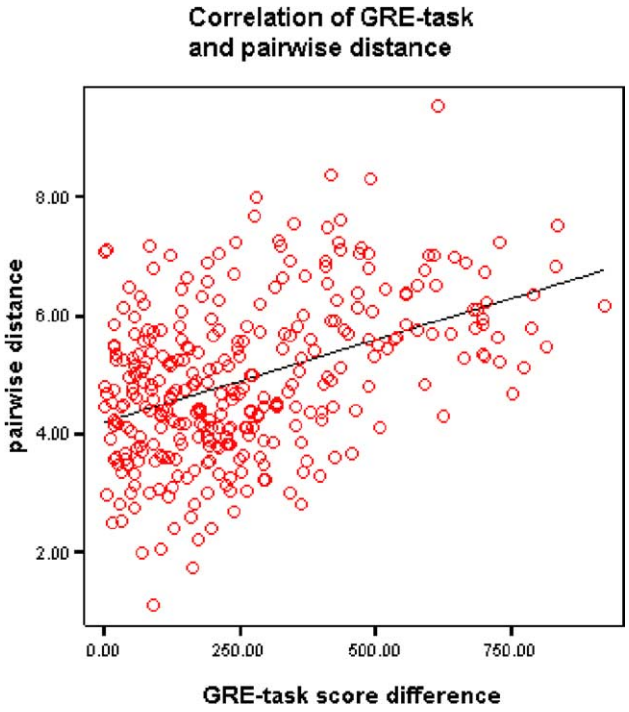

Fig. 3. Positive correlation between differences in GRE scores and differences in WSD judgments.

of subjects did not show significance as a main effect on their pair-wise disagreement. Individuals having high GRE scores, for instance, are not more likely to agree with other test subjects than are individuals with low GRE scores. Interaction *between* subjects GRE-task scores does have a significant effect on subjects' pair-wise disagreement on word senses, $F(1,3) = 8.994$, $p < 0.01$. Where both subjects have high GRE-task scores, they tend to have very low disagreement scores. However, there is more to the picture.

Where both subjects' GRE-task scores are low, the disagreement between their WSD judgments is not as great as the pair-wise disagreement between subjects with very *different* GRE-task scores. Fig. 3 shows the correlation between differences in GRE scores and differences in WSD judgments. The greatest sense disagreement between subjects is found where one subject's GRE-task score is high and the other is low. The parameter estimate for the interaction was $= -3.287\text{E}-06$, $p < 0.01$. In Fig. 4 this interaction can be seen as a bowing effect in the regression model's estimate. Not surprisingly, the same interaction is found for GRE-task scores and the pair-wise disagreement between subjects' confidence ratings, $F(1,3) = 13.681$, $p < 0.01$ (shown in Fig. 5). This may be due to a strong correlation between applicability ratings and confidence ratings. There was no significant correlation between GRE-task score and the number of sense assignments made by an individual. This finding rules out the possibility that low scoring individuals achieve higher agreement by simply assigning more senses. In fact there is a non-significant trend suggesting that individuals with high GRE-scores make more sense assignments than low scoring individuals.

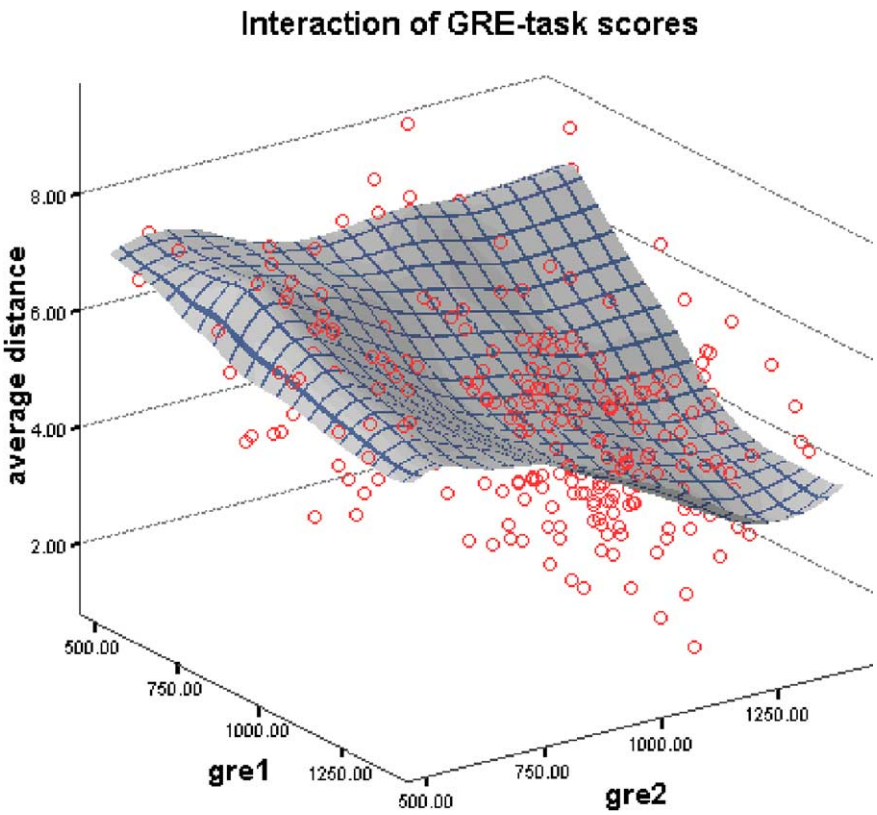

Fig. 4. Interaction of GRE task scores and differences in WSD applicability judgments.

5. Discussion

Two assumptions critical to the state of the art for word sense disambiguation have yet to be validated. The first assumption is that experts will agree more often than non-experts on lexical judgments. The second is that their level of agreement conveys some measure of "correctness". Inter-annotator agreement is problematic in this regard. The results of this study support the first assumption but challenge the second. While experts tend to agree more than non-experts do, expert knowledge alone is not the only source of agreement. If it were, we would have seen greater disparity between non-expert decisions. This study attempted to identify what some of the contributions to disagreement are by investigating the interaction of lexical ability and human judgments.

Correlation between an individual's GRE-task score and the disagreement of his WSD judgments with the gold standard (as measured by distance) suggest that there are effective measures of lexical knowledge useful to understanding performance on WSD tasks. What is interesting here is that lexical knowledge alone is not a sufficient predictor of agreement between pairs of individuals. Lexical knowledge interacts with other factors that affect agreement. This may be because individuals who know less about the language tend to make the same mistakes as others, raising

Fig. 5. Interaction of GRE task scores and differences in WSD judgment confidence.

the inter-annotator agreement without raising "correctness". It is also possible, however, that senses have different meanings to different *groups* of individuals, and that GRE-like tests capture this difference. Without consideration of the interacting factors, lexical knowledge cannot fully account for agreement or disagreement.

The relationship between time and task effectiveness in human WSD judgments is not clear. The lack of correlation between time spent on the GRE task and GRE-task scores may suggest that subjects either know the answers to lexical questions or do not (i.e., that performance scores are affected by familiarity with the vocabulary and not by careful consideration). However, subjects who rushed through the WSD-task *did* perform badly as measured against the SENSEVAL-2 gold standard. This would support the assumption that proper disambiguation requires more than just a knowledge of the vocabulary. If this is the case, sense-making strategies come into play and may even be the main cause of agreement, rather than lexical knowledge regarding vocabulary. The individuals who spent more time considering the relationships between words in the GRE task may be employing different strategies to solving lexical problems in general. Those same strategies may have enabled them to make WSD judgments that were more like those of a trained lexicographer.

The interaction effects found suggest that individuals with similar abilities will make similar judgments. We found a strong effect for lexical knowledge as measured by a standardized test.

Our results show that measures of similarity between individuals can be successfully triangulated against the larger population via standardized testing to achieve more meaningful results than those obtained from simple agreement measures.

As a practical consequence of our findings, we suggest that a group of laypersons and naïve annotators could achieve greater agreement than a group of judges with more diverse abilities. This presents a potential risk in corpus development. If an established expert is on a team of otherwise naïve judges, inter-annotator agreement will actually decrease because of the *disparity* of ability rather than increasing due to increased knowledge. The levels of agreement of those same naïve judges without the contribution of an expert would be higher. Hence, agreement alone cannot be taken as a confident measure of "correctness" but must be combined with some other measure of task ability. Defining and quantifying the appropriate level of expertise for the task is crucial to validating the task's results.

6. Conclusion

We have described and implemented a method for analyzing factors contributing to agreement between naïve judges in a word sense disambiguation task. We first addressed the assumption that expert judges would agree more often on lexical judgments than non-experts. The correlation between GRE task scores and agreement on WSD task judgments suggests that pairs of judges with greater lexical knowledge are more likely to agree than are pairs of judges with high and low levels of lexical knowledge. However, when we addressed the assumption that agreement between experts is a measure of the "correctness" of their judgments, we found significant evidence to the contrary. Correlation and regression analysis of the data indicate that pairs of non-expert judges will obtain levels of agreement comparable to pairs of expert judges. Which of these shall we tune our systems to, and how will we know which is which?

We find that a group of laypersons and naïve annotators can achieve greater agreement than a group of judges with more diverse abilities. This presents a potential risk in corpus development. If an established expert is on a team of otherwise naïve judges, inter-annotator agreement will actually decrease because of the *disparity* of ability rather than increasing due to increased expert knowledge. The levels of agreement of those same naïve judges without the contribution of an expert would be higher. Hence, agreement alone cannot be taken as a confident measure of "correctness" but must be combined with some other measure of task ability. Defining and quantifying the appropriate level of expertise for the task is crucial to validating the task's results.

The results of our study suggest that *similarity* in lexical knowledge is a strong predictor of inter-annotator agreement in word sense disambiguation even when the level of knowledge is *low*. Knowledge of the language causes subjects to converge around a meaning, but lack of knowledge does not necessarily cause two individuals to diverge. Rather, inter-annotator disagreement stems (at least partially) from *disparity* in lexical knowledge. Alternatively, if we choose to see agreement as some ill-defined form of correctness, then we must acknowledge that some of the word senses *not* chosen by experts still have a very real level of psychological correctness for naïve judges.

Our findings beg the question, if we get 95% agreement from lexical "experts" what does that agreement really mean? Is the agreement among experts due to a common understanding of how to apply lexical *tools* to a particular task, or due to an alignment of *personal* lexical knowledge?

The idea of agreement for expert judgments is tricky. Development of "gold standard" data sets should be qualified by something more than inter-annotator agreement. Validation of the operational assumptions made to date in WSD corpus development will require triangulation of knowledge and skills against more sophisticated measures.

This study has only scratched the surface of a field of investigation of potential value to WSD research. Our research calls into question the standard assumption that agreement on word sense disambiguation tasks necessarily indicates correctness of judgments. The questions we have raised about the nature of agreement require a shift of focus in the metrics of corpus quality and a refinement of estimates for the upper bound of WSD performance.

Acknowledgements

The authors would like to thank Philip Resnik and Stephanie Strassel for their input and suggestions. We thank the anonymous reviewers for their added perspective and comments. This research was funded in large part by a grant from the University of Maryland.

References

Ahlswede, T., Lorand, D., 1993. Word sense disambiguation by human subjects: Computational and psycholinguistic implications. In: Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text. 31st Annual Meeting of the Association for Computational Linguistics, pp. 1–9.

Atkins, B.T.S., Levin, B., 1991. Admitting impediments. In: Zernik, U. (Ed.), Lexical Acquisition. Lawrence Erlbaum, Hillsdale, NJ, pp. 233–262.

Cuadra, C.A., Katter, R.V., 1967. Opening the black box of "relevance". Journal of Documentation 23, 291–303.

Edmonds, P., Cotton, S., 2001. SENSEVAL-2: Overview. In: Preiss, J., Yarowsky, D. (Eds.), The Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, pp. 1–5.

Ellis, D., 1996. The dilemma of measurement in information retrieval research. Journal of the American Society for Information Science 47, 23–36.

ETS, 1997. GRE practice general test. Princeton: Educational Testing Service.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Fellbaum, C., Grabowski, J., Landes, S., 1997. Analysis of hand-tagging task. In: Palmer, M., Light, M., (Eds.), Tagging Text with Lexical Semantics: Why, What and How? pp. 34–40.

Fellbaum, C., Grabowski, J., Landes, S., 1998. Performance and confidence in a semantic annotation task. In: Fellbaum, C. (Ed.), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, pp. 217–237.

Hanks, P., 2003. Lexicography. In: Mitkov, R. (Ed.), The Oxford Handbook of Computational Linguistics. Oxford University Press, New York, pp. 48–69.

Hripcsak, G., Friedman, C., Alderson, P.O., DuMouchel, W., Johnson, S.B., Clayton, P.D., 1995. Unlocking clinical data from narrative reports: a study of natural-language processing. Annals of Internal Medicine 122, 681–688.

Jorgensen, J., 1990. The psychological reality of word senses. Journal of Psycholinguistic Research 19, 167–190.

Kilgarriff, A., 1998. Gold standard datasets for evaluating word sense disambiguation programs. Computer Speech and Language 12, 453–472.

Kilgarriff, A., Rosenzweig, J., 2000. Framework and result for English SENSEVAL. Computers and the Humanities 34, 15–48.

Krishnamurthy, R., Nicholls, D., 2000. Peeling an onion: the lexicographer's experience of manual sense-tagging. Computers and the Humanities 34, 85–97.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 140, 5–53.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J., 1990. Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3, 235–244.

Rong, T., Shaw, W., Vevea, J., 1998. Toward the identification of the optimal number of relevance categories. Journal of the American Society for Information Science 50, 254–264.

Rorvig, M.E., 1988. Psychometric measurement and information retrieval. Annual Review of Information Science and Technology 23, 157–189.

Véronis, J., 1998. A study of polysemy judgments and inter-annotator agreement. Programme and Advanced Papers of the SENSEVAL Workshop, Herstmonceux Castle (England), pp. 2–4.

Williams, J., 1992. Processing polysemous words in context: evidence for interrelated meanings. Journal of Psycholinguistic Research 21, 193–218.