Coling 2008

22nd International Conference on Computational Linguistics

Proceedings of the workshop on Human Judgements in Computational Linguistics

Workshop chairs:
Ron Artstein, Gemma Boleda,
Frank Keller, and Sabine Schulte im Walde

23 August 2008
Manchester, UK

Order copies of this and other Coling proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-905593-49-1

Design by Chimney Design, Brighton, UK
Production and manufacture by One Digital, Brighton, UK

Introduction

The papers in these proceedings were presented at the *Coling 2008 workshop on human judgements in Computational Linguistics*, held in Manchester on 23 August 2008.

Human judgements play a key role in the development and the assessment of linguistic resources and methods in Computational Linguistics. They are commonly used in the creation of lexical resources and corpus annotation, and also in the evaluation of automatic approaches to linguistic tasks: In the developmental phase, human judgements help to define an inventory of categories as well as robust annotation criteria, and in the assessment phase they are used to evaluate the results of automatic systems against existing linguistic standards. Furthermore, systematically collected human judgements provide clues for research on linguistic issues that underlie the judgement task, providing insights complementary to introspective analysis or evidence gathered from corpora.

The workshop included papers about experiments that collect human judgements for Computational Linguistic purposes. A particular focus of the workshop was concerned with human judgements on 'controversial' linguistic tasks (those that are not clear from a theoretical point of view, such as many tasks having to do with semantics or pragmatics). Such experimental tasks are usually difficult to design and interpret, and they typically result in low agreement scores. They are often poorly documented in the literature; however, they are especially well-suited as a basis for a fruitful discussion.

We were pleased to receive 22 submissions, of which we accepted eight papers for presentation at the workshop. We would like to thank the program committee for the excellent job they did in reviewing the submissions. As well as the paper presentations, the workshop also included a keynote talk by Martha Palmer of the University of Colorado at Boulder. These proceedings include an abstract of her talk.

The workshop was supported by the Spanish Education and Science Ministry via the KNOW project (TIN2006-15049-C03-03) and by Sonderforschungsbereich 732, Universität Stuttgart.

Ron Artstein
Gemma Boleda
Frank Keller
Sabine Schulte im Walde

Organizers:

Ron Artstein, University of Southern California
Gemma Boleda, Universitat Politècnica de Catalunya
Frank Keller, University of Edinburgh
Sabine Schulte im Walde, Unversität Stuttgart

Programme Committee:

Toni Badia, Universitat Pompeu Fabra
Marco Baroni, University of Trento
Beata Beigman Klebanov, Northwestern University
André Blessing, Universitä Stuttgart
Chris Brew, Ohio State University
Kevin Cohen, University of Colorado Health Sciences Center
Barbara Di Eugenio, University of Illinois at Chicago
Katrin Erk, University of Texas at Austin
Stefan Evert, University of Osnabrück
Afsaneh Fazly, University of Toronto
Alex Fraser, Universität Stuttgart
Jesus Gimenez, Universitat Politècnica de Catalunya
Roxana Girju, University of Illinois at Urbana-Champaign
Ed Hovy, University of Southern California
Nancy Ide, Vassar College
Adam Kilgarriff, University of Brighton
Alexander Koller, University of Edinburgh
Anna Korhonen, University of Cambridge
Mirella Lapata, University of Edinburgh
Diana McCarthy, University of Sussex
Alissa Melinger, University of Dundee
Paola Merlo, University of Geneva
Sebastian Padó, Stanford University
Martha Palmer, University of Colorado
Rebecca Passonneau, Columbia University
Massimo Poesio, University of Trento
Sameer Pradhan, BBN Technologies
Horacio Rodriguez, Universitat Politècnica de Catalunya
Bettina Schrader, Universität Potsdam
Suzanne Stevenson, University of Toronto

Invited Speaker:

Martha Palmer, University of Colorado at Boulder

Table of Contents

Invited Talk: *The Relevance of a Cognitive Model of the Mental Lexicon to Automatic Word Sense Disambiguation*
Martha Palmer and Susan Brown 1

Analyzing Disagreements
Beata Beigman Klebanov, Eyal Beigman and Daniel Diermeier 2

Exploiting 'Subjective' Annotations
Dennis Reidsma and Rieks Op den Akker 8

Human Judgement as a Parameter in Evaluation Campaigns
Jean-Baptiste Berthelin, Cyril Grouin, Martine Hurault-Plantet and Patrick Paroubek 17

Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection
Joel Tetreault and Martin Chodorow 24

Polysemy in Verbs: Systematic Relations between Senses and their Effect on Annotation
Anna Rumshisky and Olga Batiukova 33

Eliciting Subjectivity and Polarity Judgements on Word Senses
Fangzhong Su and Katja Markert 42

Human Judgements in Parallel Treebank Alignment
Martin Volk, Torsten Marek and Yvonne Samuelsson 51

An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text
Plaban Kumar Bhowmick, Anupam Basu and Pabitra Mitra 58

Workshop Programme

Saturday, 23 August 2008

9:20–9:30	Introductory remarks
9:30–10:30	Invited Talk: *The Relevance of a Cognitive Model of the Mental Lexicon to Automatic Word Sense Disambiguation* Martha Palmer and Susan Brown
10:30–11:00	Break
11:00–11:30	*Analyzing Disagreements* Beata Beigman Klebanov, Eyal Beigman and Daniel Diermeier
11:30–12:00	*Exploiting 'Subjective' Annotations* Dennis Reidsma and Rieks Op den Akker
12:00–12:30	*Human Judgement as a Parameter in Evaluation Campaigns* Jean-Baptiste Berthelin, Cyril Grouin, Martine Hurault-Plantet and Patrick Paroubek
12:30–14:00	Lunch
14:00–14:30	*Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection* Joel Tetreault and Martin Chodorow
14:30–15:00	*Polysemy in Verbs: Systematic Relations between Senses and their Effect on Annotation* Anna Rumshisky and Olga Batiukova
15:00–15:30	*Eliciting Subjectivity and Polarity Judgements on Word Senses* Fangzhong Su and Katja Markert
15:30–16:00	Break
16:00–16:30	*Human Judgements in Parallel Treebank Alignment* Martin Volk, Torsten Marek and Yvonne Samuelsson
16:30–17:00	*An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text* Plaban Kumar Bhowmick, Anupam Basu and Pabitra Mitra
17:00–17:30	Discussion

The Relevance of a Cognitive Model of the Mental Lexicon to Automatic Word Sense Disambiguation

Martha Palmer and Susan Brown
University of Colorado at Boulder
Department of Linguistics
Hellems 290, 295 UCB
Boulder, CO 80309-0295, USA

Supervised word sense disambiguation requires training corpora that have been tagged with word senses, and these word senses typically come from a pre-existing sense inventory. Space limitations imposed by dictionary publishers have biased the field towards lists of discrete senses for an individual lexeme. Although some dictionaries use hierarchical entries to emphasize relations between senses, many do not. WordNet, which has been the default choice of NLP researchers for sense tagging because of its broad coverage and easy accibility, does not have hierarchical entries. Could the relations between senses that are captured by a hierarchy be useful to NLP systems? Concerns have also been raised about whether or not WordNet's word senses are unnecessarily fine-grained. WSD systems are obviously more successful in distinguishing coarse-grained senses than fine-grained ones (Navigli, 2006), but important information could be lost if fine-grained distinctions are ignored. Recent psycholinguistic evidence seems to indicate that closely related word senses may be represented in the mental lexicon much like a single sense, whereas distantly related senses may be represented more like discrete entities (Brown, 2008). These results suggest that, for the purposes of WSD, closely related word senses can be clustered together into a more general sense with little meaning loss. This talk will describe this psycholinguistic research and its current implications for automatic word sense disambiguation, as well as plans for future research and its possible impact.

Analyzing Disagreements

Beata Beigman Klebanov, Eyal Beigman, Daniel Diermeier
Kellogg School of Business
Northwestern University
{beata,e-beigman,d-diermeier}@northwestern.edu

Abstract

We address the problem of distinguishing between two sources of disagreement in annotations: genuine subjectivity and slip of attention. The latter is especially likely when the classification task has a default class, as in tasks where annotators need to find instances of the phenomenon of interest, such as in a metaphor detection task discussed here. We apply and extend a data analysis technique proposed by Beigman Klebanov and Shamir (2006) to first distill reliably deliberate (non-chance) annotations and then to estimate the amount of attention slips vs genuine disagreement in the reliably deliberate annotations.

1 Introduction

Classification tasks fall into two broad categories. Those in the first category proceed by requiring that every item is explicitly assigned a tag out of a given set of tags; part-of-speech tagging is an example (Santorini, 1990).

In the second group of tasks, the annotator is asked to identify a phenomenon of interest, thus implicitly classifying items as belonging to the phenomenon (marked) and not belonging to it (left unmarked). When the studied phenomenon is expected to have low incidence, this is a time-saving strategy, as annotators do not need to bother with explicitly marking (almost) everything as a non-phenomenon. A recent example of such a task is Beigman Klebanov and Shamir (2006), where annotators were asked to provide anchors for words deemed anchored in the text (i.e. associatively connected to a previous item in the text), thus leaving words that did not receive an anchor implicitly marked as un-anchored. Psychological experiments where people are asked to respond to the occurrence of a given phenomenon can also be viewed as implicit classifications; for example, see Spiro's (2007) work on identification of boundaries of musical phrases by listeners. The task of metaphor detection discussed in this paper also falls under the implicit classification category.

While such a strategy uses annotators' time efficiently, some of the observed disagreements could be due to an annotator missing an occurrence of the relevant phenomenon, rather than genuinely disagreeing on the matter of occurrence.

We show in section 2 that our metaphor identification task features less-than-perfect inter-annotator agreement. Section 3 uses Beigman Klebanov and Shamir's (2006) methodology to find annotations that can be reliably attributed to a deliberate decision by at least some of the annotators. We then discuss the use of validation experiment to distinguish between slips of attention and genuine disagreements (sections 4,5).

2 Metaphor Detection Study

For a project studying the use of metaphors in public discourse, a dataset of 151 articles from the British press was subjected to annotation.[1] Participants were asked to mark paragraphs that contain occurrences of metaphors from LOVE, VEHICLE, AUTHORITY and BUILDING domains (henceforth, **metaphor types**).

For example, the following paragraph in 20 September 1992 issue of *Sunday Times* contains an

[1] This is part of the data discussed in (Musolff, 2000).

extended metaphor from the VEHICLE domain:

> Thatcher warned EC leaders to stop their endless round of summits and take notice of their own people. “There is a fear that the European train will thunder forward, laden with its customary cargo of gravy, towards a destination neither wished for nor understood by electorates. But the train can be stopped,” she said.

The title[2] of one of the articles in the 19 October 1999 issue of *The Guardian* contains a LOVE metaphor:

> Euro-flirting is not only a matter of desire.

The discussion in this paper is based on the output of 9 annotators who performed metaphor identification (henceforth, **production task**), and of 7 annotators (out of 9) who took part in the subsequent validation study (henceforth, **validation task**). Subjects were not told about validation until after they finished production on the whole of the dataset. A time gap of 2 weeks existed between the end of the production study and the start of the validation, each of the tasks taking 6 weeks, in weekly installments of 25 articles each.

For the production task, the annotators were instructed to mark every paragraph where a metaphor from the given metaphor type appeared; the 151-article dataset yields 2364 paragraphs. This paradigm corresponds to the implicit classification task discussed earlier, in that only the positive (metaphor-containing) cases are given an explicit markup. The incidence of positive cases is quite low – VEHICLE, the most ubiquitous type, featured in 4% of the paragraphs, on average across annotators.

We note that the appearances of the different metaphor types are not mutually exclusive, and, indeed, there is no a-priori reason to suppose any relationship between them. For example, the following paragraph from the leading article in 15 November 1995 issue of *The Guardian* was marked by some annotators as containing both LOVE and VEHICLE metaphors:

> The first European bank notes - probably to be called “euros” - will not be in circulation until 2002 judging by yesterday’s report from the European Monetary Institute. But this doesn’t mean that monetary union has been delayed beyond 1999 because the printing of European bank notes will have been preceded by a period of three years when national currencies will have been locked together in indissoluble monetary matrimony [...] Although France looks as if it might buckle under the strain of meeting the fiscal criteria and in Germany the SDP is having doubts (though only about whether the new currency will be strong enough) the Maastricht train is still theoretically on the rails. Nobody has changed the timetable.

We therefore treat the detection of metaphors from each metaphor type as a separate binary classification task. Table 1 shows the inter-annotator agreement for the production task using the κ statistic (Carletta, 1996; Krippendorff, 1980; Siegel and Castellan, 1988).

Table 1: Metaphor annotation data (production), by metaphor type. The third column shows the percentage of paragraphs (out of 2364) marked as having a metaphor of the given type, on average across 9 annotators.

Type	κ	marked
VEHICLE	0.66	4.0%
LOVE	0.66	2.5%
AUTHORITY	0.39	2.7%
BUILD	0.43	1.7%

Clearly, it is not the case that the whole of the dataset was reliably annotated, even for the better-agreed-upon metaphor types like VEHICLE and LOVE. Hence, additional procedures are needed to distill reliable annotations. We apply Beigman Klebanov and Shamir’s (2006) statistical technique to find a subset of the data that is sufficiently reliable, and later corroborate the statistical analysis through the validation task.

3 Reliably Deliberate Annotations

In Beigman Klebanov and Shamir (2006), 22 subjects performed the anchoring annotation; the overall inter-annotator agreement was κ=0.45.

[2] A title is treated as a paragraph in our annotations.

Thus, some of the data was clearly unreliable, as in our metaphor detection task, but the possibility existed that some other part was in fact annotated sufficiently reliably.

Beigman Klebanov and Shamir's (2006) analysis proceeded thus: Suppose each of the 20 annotators[3] ($i = 1...20$) was flipping a coin with the probability of heads p_i equal to the proportion of "anchored" markups in annotator i's data. What is the level of agreement for which this scenario is sufficiently improbable? For their data, the random anchoring hypothesis could be rejected with 99% confidence for cases marked by at least 13 people. Items featuring at least this level of agreement can be considered, with high probability, as **deliberately annotated** as "anchored", as at least some of those who marked them were not flipping a coin.

Following the procedure in Beigman Klebanov and Shamir (2006), we wish to determine a reliably deliberate subset of our metaphor annotations. We induce 9 random pseudo-annotators from the 9 actual ones, each marking paragraphs at random as containing a metaphor of a given type or not. Pseudo-annotator i flips a coin with $p(heads) = p_i$, which is the proportion of metaphor markups by the i'th annotator for the most common metaphor type (VEHICLE).

Assuming each annotator flips her coin, we calculate the probability of 3 or more coins coming up heads simultaneously;[4] this probability is 0.0045. Thus, with 99.5% confidence, a metaphor markup by at least 3 people is not a result of coinflip, at least for some of the annotators. We note, however, that 99.5% confidence is insufficient for our case: It allows for random highly agreed markup in 0.5% of the instances. Given that only up to 4% of the instances have positive markups, this would yield a high percentage of random items in the positive instances. The probability of 4 or more pseudo-annotators having their coins come up heads simultaneously is below 0.0003; we consider this sufficient confidence for our case, and regard metaphor markups produced by at least 4 people as reliably deliberate.

Note that we cannot find a similar threshold for no-metaphor annotations, as a lack of metaphor annotation could happen by chance with a high probability ($p = 0.69$). In view of the potential use of the dataset for evaluating metaphor detection algorithms, a putative metaphor suggested by the algorithm cannot be rejected based on the lack of metaphor annotation in the data. A complementary procedure would be needed, for example, collecting human judgments for the putative metaphors separately.

4 Attention Slips vs Genuine Disagreements

Deliberate annotation does not guarantee agreement. It remained the case that some of the reliably deliberate data in Beigman Klebanov and Shamir (2006) was actually produced by only some of the original subjects. Indeed, some of the deliberately marked metaphors were annotated by only 4 out of the 9 participants. For cases where the positive annotations were produced deliberately, what is the status of negative annotations accorded to the same items? Were these mere attention slips, or genuine differences of opinion? Note that this question cannot be meaningfully posed regarding the parts of annotations for which the hypothesis of random positive marking could not be rejected with sufficiently high probability, since, obviously, apparent disagreements there could be simply a result of different coinflip outcomes.

Beigman Klebanov and Shamir (2006) hypothesized that dissenting annotations of the reliable pairs would be cases of attention slips, rather than genuine differences of opinion. In other words, while there was no initial agreement, these items were potentially *agreeable*. To test the hypothesis, they devised a validation experiment, where subjects were presented with all pairs marked by at least one annotator, plus some random pairs, and were asked to cross out things they disagree with. The reasoning was as follows: If attention slip was the cause for a dissenting negative annotation, when the subject is asked about the relevant item, i.e. it is explicitly brought to her attention, she would accept it, whereas if a case is that of a genuine disagreement, she would reject it. To control for the possibility that people just accept everything so that not to be dissonant with others, some random annotations were also included.

The results reported by Beigman Klebanov and Shamir (2006) largely bore out the hypothesis. First, people did not tend to accept everything,

[3] Two people were excluded as outliers.

[4] In Beigman Klebanov and Shamir (2006), a normal approximation is used to handle collective decision making by 20 pseudo-annotators. In the current case, 9 annotators is a sufficiently small number to allow an exact probability calculation over the 512 possibilities.

as only 15% of judgments of random annotations and only 62% of judgments on all human-generated annotations were "accept" judgments. However, 94% of judgments of the reliable annotations were "accept" judgments. Hence, the rate of genuine disagreement on the reliably deliberate part of Beigman Klebanov and Shamir's (2006) data turned out to be quite low.

We are interested in estimating the degree of genuine disagreements in metaphor production. Using Beigman Klebanov and Shamir's methodology, we collected all paragraphs marked as containing a metaphor of a given type by at least one of the 9 annotators, plus added random markups. This data was submitted to 7 subjects for validation.

Table 2: Percentage of "Accept" validations for random (Rand) and human (Hum) metaphor production data, as well as for the partition of the human data into reliably deliberate (Rel) and unreliable (URel) subsets. For each subset, the number of data instances covered by the subset is shown. Subscripts indicate metaphor type: (V)EHICLE, (L)OVE, (A)UTHORITY, (B)UILD. The bottom line shows the average over metaphor types.

Subset	#	Acc	Subset	#	Acc
Rand$_V$	94	5%	Hum$_V$	194	73%
Rand$_L$	56	6%	Hum$_L$	137	64%
Rand$_A$	62	12%	Hum$_A$	258	51%
Rand$_B$	40	1%	Hum$_B$	126	68%
Rand	252	6%	Hum	715	62%

Subset	#	Acc	Subset	#	Acc
URel$_V$	92	49%	Rel$_V$	102	94%
URel$_L$	81	43%	Rel$_L$	56	95%
URel$_A$	218	42%	Rel$_A$	40	96%
URel$_B$	86	55%	Rel$_B$	40	96%
URel	477	46%	Rel	238	95%

Table 2 reports the percentage of "accept" votes for random and human metaphor production data, as well as for reliably deliberate and unreliable subsets of the human data. As in Beigman Klebanov and Shamir's case, the validation experiment clearly distinguishes between random, human in general, and reliably deliberate subsets, and puts the estimated degree of genuine disagreement in metaphor identification at 5% on average, with little variation across the metaphor types. That is, given that, with high probability, at least some humans deliberately identified a paragraph as containing a metaphor, the chance for its rejection is about 5%. The rest of observed production disagreements, for the reliably deliberate subset, are remedied at validation time, thus probably constituting attention slips during production. The reliably deliberate subset contains 33% (238/715) of all human-generated data.

5 Separating self and others

One potential confounder in the above analysis is conflation of self-consistency with affirmation of someone else's annotations. It is possible that many of the validation-time "accept" votes are cases of people accepting their own earlier annotation; the proportion of such cases is expected to increase the more people marked the metaphor during production. Therefore, to get a more precise estimate of the degree of genuine disagreement, we control for self-affirmation, and calculate the proportion of "accept" validations in cases where the person did not mark the metaphor during production. Specifically, if X of the 7 people who participated in both production and validation marked the metaphor at production,[5] we check the split of the remaining 7-X votes during validation. Table 3 presents average other-affirmation rates for the reliably deliberate and unreliable human produced data. Note that only 184 out of the 238 deliberately reliable cases can be used, as the remaining 54 are cases where all 7 annotators produced the markup, so there is no disagreement.

Table 3: Percentage of "Accept" validations for reliably deliberate (Rel) and unreliable (URel) subsets of the metaphor production data, given that the subject himself did NOT produce the metaphor.

Subset	#	Acc	Subset	#	Acc
URel$_V$	92	44%	Rel$_V$	78	90%
URel$_L$	81	39%	Rel$_L$	38	92%
URel$_A$	218	35%	Rel$_A$	30	91%
URel$_B$	86	53%	Rel$_B$	38	91%
URel	477	41%	Rel	184	91%

[5] The actual total of the production annotations could be up to X+2, as there were 2 more annotators in production than in validation.

According to the table, 91% cases of disagreements in the reliably deliberate data are remedied at validation time. That is, given that, with high probability, at least some human deliberately identified a paragraph as containing a metaphor, the chance for its rejection by *a human who initially apparently disagreed with the annotation* is only about 9%.

Finally, validation data allows an investigation of the stability of people's judgments by calculating self-rejection rates, i.e. estimating the probability of rejecting during validation an instance that the same annotator marked as containing a metaphor during production. Table 4 shows the results.

Table 4: Percentage of "Reject" validations for reliably deliberate (Rel) and unreliable (URel) subsets of the metaphor production data, given that the subject himself produced the annotation.

Subset	#	Rej	Subset	#	Rej
URel$_V$	72	25%	Rel$_V$	102	4%
URel$_L$	55	26%	Rel$_L$	56	5%
URel$_A$	198	22%	Rel$_A$	40	2%
URel$_B$	60	23%	Rel$_B$	40	2%
URel	385[6]	23%	Rel	238	4%

For the reliably deliberate data, i.e. cases where at least 4 people produced the markup, the average self-rejection rate is 4%. This low figure further supports the designation of the reliably deliberate subset as such, i.e. containing stable annotations, as in 96% of the cases a person who produced the markup is likely to re-affirm it when asked again, even after a substantial time delay.[7]

For the "unreliable" data, i.e. cases where only one or two people marked the metaphor during production, the average self-rejection rate is 23%. Self-rejection means either that the initial positive markup was a mistake, or that it is difficult for the annotator to make up his mind about the annotation of the item. In any case, high self-rejection rate means that the relevant production annotations cannot be trusted to contain a settled judgment that could be then agreed or disagreed with by other annotators, or indeed replicated by a computational model.

We consider self-rejected cases potential indicators of a difficulty on the annotator's part to decide on the correct markup. We plan a more detailed investigation of the materials to see whether these cases exhibit any interesting common properties that could help characterize the difficulties in metaphor identification task.

6 Conclusion

In this article, we showed an application of Beigman Klebanov and Shamir's (2006) methodology for analyzing annotation data to metaphor identification annotations. The approach allowed establishing an agreement threshold beyond which the annotations are reliably deliberate, in the sense that, with high probability, at least some of the annotators who detected a metaphor were not flipping a coin. This threshold is agreement of 4 out of 9 annotators, for 99.9% reliability.

To investigate the nature of disagreements in the reliably deliberate subset, we followed Beigman Klebanov and Shamir (2006) in conducting a validation study, where subjects were asked to accept or reject markups produced during the initial annotation study, as well as some random annotations. Sharpening the methodology somewhat, we showed that in 91% of reliably deliberate cases where an annotator did not produce the markup himself, he accepted it during validation. Hence, the bulk of the initial disagreements were amended during validation, with the residual 9% being likely locations for genuine difference of opinion.

Further analysis of validation data revealed that the reliably deliberate subset features low self-rejection rates, meaning that people are consistent with their own production. This was not the case for the subset deemed unreliable during statistical analysis, where a 23% self-rejection rate was observed. We hypothesize that some of these would be hard-to-decide cases with respect to the metaphor identification task, and hence warrant a closer look in order to characterize annotator difficulties with the task.

[6] Note that only 385 of the 477 items in the unreliable data could be used for the calculation. The remaining items were not produced by any of the 7 people who participated in both production and validation, but only by one or both of the 2 additional production-task annotators.

[7] The time difference between production and validation per article ranged between 4 and 8 weeks, due to differences in the order in which the different subjects were given the articles.

7 Acknowledgment

We would like to thank an anonymous reviewer for a thorough and insightful review that helped as improve this article significantly.

References

Beigman Klebanov, Beata and Eli Shamir. 2006. Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 40(2):109–126.

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

Krippendorff, Klaus. 1980. *Content Analysis*. Sage Publications.

Musolff, Andreas. 2000. *Mirror images of Europe: Metaphors in the public debate about Europe in Britain and Germany*. München: Iudicium.

Santorini, Beatrice. 1990. Part-of-speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

Siegel, Sidney and John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Book Company.

Spiro, Neta. 2007. *What contributes to the perception of musical phrases in Western classical music?* Ph.D. thesis, University of Amsterdam, The Netherlands.

Exploiting ‘Subjective’ Annotations

Dennis Reidsma
Human Media Interaction
University of Twente, PO Box 217
NL-7500 AE, Enschede, The Netherlands
dennisr@ewi.utwente.nl

Rieks op den Akker
Human Media Interaction
University of Twente, PO Box 217
NL-7500 AE, Enschede, The Netherlands
infrieks@ewi.utwente.nl

Abstract

Many interesting phenomena in conversation can only be annotated as a subjective task, requiring interpretative judgements from annotators. This leads to data which is annotated with lower levels of agreement not only due to errors in the annotation, but also due to the differences in how annotators interpret conversations. This paper constitutes an attempt to find out how subjective annotations with a low level of agreement can profitably be used for machine learning purposes. We analyse the (dis)agreements between annotators for two different cases in a multimodal annotated corpus and explicitly relate the results to the way machine-learning algorithms perform on the annotated data. Finally we present two new concepts, namely ‘subjective entity’ classifiers resp. ‘consensus objective’ classifiers, and give recommendations for using subjective data in machine-learning applications.

1 Introduction

Research that makes use of multimodal annotated corpora is always presented with something of a dilemma. One would prefer to have results which are reproducible and independent of the particular annotators that produced the corpus. One needs data which is annotated with as few disagreements between annotators as possible. But labeling a corpus is a task which involves a judgement by the annotator and is therefore, in a sense, always a subjective task. Of course, for some phenomena those judgements can be expected to come out mostly the same for different annotators. For other phenomena the judgements can be more dependent on the annotator *interpreting* the behavior being annotated, leading to annotations which are more subjective in nature. The amount of overlap or agreement between annotations is then also influenced by the amount of *intersubjectivity* in the judgements of annotators.

This relates to the spectrum of content types discussed extensively by Potter and Levine-Donnerstein (1999). One of the major distinctions that they make is a distinction in annotation of *manifest content* (directly observable events), *pattern latent content* (events that need to be inferred indirectly from the observations), and *projective latent content* (loosely said, events that require a subjective interpretation from the annotator).

Manifest content is what is directly observable. Some examples are annotation of instances where somebody raises his hand or raises an eyebrow, annotation of the words being said and indicating whether there is a person in view of the camera. Annotating manifest content can be a relatively easy task. Although the annotation task involves a judgement by the annotator, those judgements should not diverge a lot for different annotators.

At the other end of the spectrum we find *projective latent content*. This is a type of content for which the annotation schema does not specify in extreme detail the rules and surface forms that determine the applicability of classes, but in which the coding relies on the annotators’ existing mental conception[1] of the classes. Such an ap-

[1] Potter and Levine-Donnerstein use the word “mental scheme” for this. We will use “mental conceptions” in this

proach is useful for everyday concepts that most people understand and to a certain extent share a common meaning for, but for which it is almost impossible to provide adequately complete definitions. Potter and Levine-Donnerstein use the example 'chair' for everyday concepts that are difficult to define exhaustively. But this concept is also especially relevant in an application context that *requires the end user of the data to agree with the distinctions being made*. This is very important when machine learning classifiers are developed to be used in everyday applications. For example, one can make a highly circumscribed, ethologically founded definition of the class 'dominant' to guide annotation. This is good for, e.g., research into social processes in multiparty conversations. However, in a scenario where an automatic classifier, trained to recognize this class, is to be used in an application that gives a participant in a meeting a quiet warning when he is being too dominant (Rienks, 2007) one would instead prefer the class rather to fit the mental conceptions of dominance that a 'naive' user may have. When one designs an annotation scheme for projective latent content, the focus of the annotation guidelines is on instructions that trigger the appropriate existing mental conceptions of the annotators rather than on writing exhaustive descriptions of how classes can be distinguished from each other (Potter and Levine-Donnerstein, 1999).

Interannotator agreement takes on different roles for the two ends of the spectrum. For manifest content the level of agreement tells you something about how accurate the measurement instrument (schema plus coders) is. Bakeman and Gottman, in their text book *observing interaction: introduction to sequential analysis* (1986, p 57), say about this type of reliability measurement that it is a matter of "calibrating your observers". For projective content, we have additional problems; the level of agreement may be influenced by the level of intersubjectivity, too. Where Krippendorff (1980) describes that annotators should be interchangeable, annotations of projective latent content can sometimes say as much about the mental conceptions of the particular annotator as about the person whose interactions are being annotated. The personal interpretations of the data by the annotator should not necessarily be seen as 'errors', though, even if those interpretations lead to low interannotator agreement: they may simply be an unavoidable aspect of the interesting type of data one works with.

Many different sources of low agreement levels, and many different solutions, are discussed in the literature. It is important to note that some types of disagreement are more systematic and other types are more noise like. For projective latent content one would expect more consistent *structure* in the disagreements between annotators as they are caused by the differences in the personal ways of interpreting multimodal interaction. Such systematic disagreements are particularly problematic for subsequent use of the data, more so than noise-like disagreements. Therefore, an analysis of the quality of an annotated corpus should not stop at presenting the value of a reliability metric; instead one should investigate the patterns in the disagreements and discuss the possible impact they have on the envisioned uses of the data (Reidsma and Carletta, 2008). Some sources of disagreements are the following.

(1) *'Clerical errors'* caused by a limited view of the interactions being annotated (low quality video, no audio, occlusions, etc) or by slipshod work of the annotator or the annotator misunderstanding the instructions. Some solutions are to provide better instructions and training, using only good annotators, and using high quality recordings of the interaction being annotated.

(2) *'Invalid or imprecise annotation schemas'* that contain classes that are not relevant or do not contain classes that are relevant, or force the annotator to make choices that are not appropriate to the data (e.g. to choose one label for a unit where more labels are applicable). Solutions concern redesigning the annotation schema, for example by merging difference classes, allowing annotators to use multiple labels, removing classes, or adding new classes.

(3) *'Genuinely ambiguous expressions'* as described by Poesio and Artstein (2005). They discuss that disagreements caused by ambiguity are not so easily solved.

(4) *'A low level of intersubjectivity'* for the interpretative judgements of the annotators, caused by the fact that there is less than perfect overlap between the mental conceptions of the annotators. The solutions mentioned above for issue (2) partly also apply here. However, in this article we focus on an additional, entirely different, way of coping

paper to avoid confusion with the term "annotation scheme".

with disagreements resulting from a low level of intersubjectivity that actively exploits the systematic differences in the annotations caused by this.

1.1 Useful results from data with low agreement

Data with a low interannotator agreement may be difficult to use, but there are other fields where partial solutions have been found to the problem, such as the information retrieval evaluation conferences (TREC). Relevance judgements in TREC assessments (and document relevance in general) are quite subjective and it is well known that agreement for relevance judgements is not very high (Voorhees and Harman report 70% three-way percent agreement on 15,000 documents for three assessors (1997)). Quite early in the history of the TREC, Voorhees investigated what the consequences of this low level of agreement are for the usefulness of results obtained on the TREC collection. It turns out that specifying a few constraints[2] is enough to be able to use the TREC assessments to obtain meaningful evaluation results (Voorhees, 2000). Inspired by this we try to find ways of looking at subjective data that tells us what constraints and restrictions on the use of it follow from the patterns in the disagreements between annotators, as also advised by Reidsma and Carletta (2008).

1.2 Related Work

In corpus research there is much work with annotations that need subjective judgements of a more subjective nature from an annotator about the behavior being annotated. This holds for Human Computer Interaction topics such as affective computing or the development of Embodied Conversational Agents with a personality, but also for work in computational linguistics on topics such as emotion (Craggs and McGee Wood, 2005), subjectivity (Wiebe et al., 1999; Wilson, 2008) and agreement and disagreement (Galley et al., 2004).

If we want to interpret the results of classifiers in terms of the patterns of (dis)agreement found between annotators, we need to subject the classifiers with respect to each other and to the 'ground truth data' to the same analyses used to evaluate and compare annotators to each other. Vieira (2002) and Steidl et al. (2005) similarily remark that it is not 'fair' to penalize machine learning performance for errors made in situations where humans would not agree either. Vieira however only looks at the *amount* of disagreement and does not explicitly relate the classes where the system and coders disagree to the classes where the coders disagree with each other. Steidl et al.'s approach is geared to data which is multiply coded for the whole corpus (very expensive) and for annotations that can be seen as 'additive', i.e., where judgements are not mutually exclusive.

Passonneau et al. (2008) present an extensive analysis of the relation between per-class machine learning performance and interannotator agreement obtained on the task of labelling text fragments with their function in the larger text. They show that overall high agreement can indicate a high learnability of a class in a multiply annotated corpus, but that the interannotator agreement is not necessarily predictive of the learnability of a label from a single annotator's data, especially in the context of what we call projective latent content.

1.3 This Paper

This paper constitutes an attempt to find out how subjective annotations, annotated with a low level of agreement, can profitably be used for machine learning purposes. First we present the relevant parts of the corpus. Subsequently, we analyse the (dis)agreements between annotators, on more aspects than just the value of a reliability metric, and explicitly relate the results to the way machine-learning algorithms perform on the annotated data. Finally we present two new concepts that can be used to explain and exploit this relation ('subjective entity' classifiers resp. 'consensus objective' classifiers) and give some recommendations for using subjective data in machine-learning applications.

2 From Agreement to Machine Learning Performance

We used the hand annotated face-to-face conversations from the 100 hour AMI meeting corpus (Carletta, 2007). In the scenario-based AMI meetings, design project groups of four players have the task to design a new remote TV control. Group members have roles: project manager (PM), industrial designer (ID), user interface design (UD), and marketing expert (ME). Every group has four meetings (20-40 min. each), dedicated to a subtask. Most of

[2] Only discuss *relative* performance differences on different (variations of) algorithms/systems run on *exactly the same set of assessments* using the *same set of topics*.

the time the participants sit at a square table.

The meetings were recorded in a meeting room stuffed with audio and video recording devices, so that close facial views and overview video, as well as high quality audio is available. Speech was transcribed manually, and words were time aligned. The corpus has several layers of annotation for several modalities, such as dialogue acts, topics, hand gestures, head gestures, subjectivity, visual focus of attention (FOA), decision points, and summaries, and is easily extendible with new layers. The *dialogue act* (DA) layer segments speaker turns into dialogue act segments, on top of the word layer, and they are labeled with one of 15 dialogue act type labels, following an annotation procedure.

In this section we will inspect (dis)agreements and machine learning performance for two corpus annotation layers: the addressing annotations (Jovanović et al., 2006) and for a particular type of utterances in the corpus, the *"Yeah-utterances"* (Heylen and op den Akker, 2007).

2.1 Contextual Addressing

A part of the AMI corpus is also annotated with addressee information. Real dialogue acts (i.e. all dialogue acts but backchannels, stalls and fragments) were assigned a label indicating who the speaker addresses his speech to (is talking to). In these type of meetings most of the time the speaker addresses the whole group, but sometimes his dialogue act is particularly addressed to some individual (about 2743 of the 6590 annotated real dialogue acts); for example because he wants to know that individual's opinion. The basis of the concept of addressing underlying the addressee annotation in the AMI corpus originates from Goffman (Goffman, 1981). The addressee is the participant *"oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants"*. Sub-group addressing hardly occurs and was not annotated. Thus, DAs are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*) (see Jovanovic et al. (2006)).

Another layer of the corpus contains *focus of attention* information derived from head, body and gaze observations (Ba and Odobez, 2006), so that for any moment it is known whether a person is looking at the table, white board, or some other participant. Gaze and focus of attention are important elements of addressing behavior, and therefore FOA is a strong cue for the annotator who needs to determine the addressee of an utterance. However, FOA is not the only cue. Other relevant cues are, for example, proper names and the use of addressing terms such as "you". Even when the gaze is drawn to a projection screen, or the meeting is held as a telephone conference without visuals, people are able to make the addressee of their utterances clear.

From an extensive (dis)agreement analysis of the addressing and FOA layers the following conclusions can be summarized: the visual focus of attention was annotated with a very high level of agreement (Jovanović, 2007); in the addressee annotation there is a large confusion between DAs being G-addressed or I-addressed; if the annotators agree on an utterance being I-addressed they typically also agree on the particular individual being addressed; 'elicit' DAs were easier to annotate with addressee than other types of dialog act; and reliability of addressee annotation is dependent on the FOA context (Reidsma et al., 2008). When the speaker's FOA is not directed to any participant the annotators must rely on other cues to determine the addressee and will disagree a lot more than when they are helped by FOA related cues. Some of these disagreements can be due to systematic subjective differences, e.g. an annotator being biased towards the 'Group' label for utterances that are answers to some question. Other disagreements may be caused by the annotator being forced to choose an addressee label for utterances that were not be clearly addressed in the first place.

In this section we will not so much focus on the *subjectivity* of the addressee annotation as on the *multimodal context* in which annotators agree more. Specifically, we will look further at the way the level of agreement with which addressee has been annotated is dependent on the FOA context of a set of utterances. We expect this will be reflected directly by the machine learning performance in these two contexts: the low agreement might indicate a context where addressee is inherently difficult to determine and furthermore the context with high agreement will result in annotations containing more consistent information that machine learning can model.

To verify this assumption we experimented with automatic detection of the addressee of an utterance based on lexical and multimodal features. Compared to Jovanovič (2007), we use a limited set of features that does not contain local context features such as ‘previous addressee’ or ‘previous dialogue act type’. Besides several lexical features we also used features for focus of attention of the speaker and listeners during the utterance. Below we describe two experiments with this task. Roughly 1 out of every 3 utterances is performed in a context where the speaker’s FOA is not directed at any other participant. This gives us three contexts to train and to test on: all utterances, all utterances where the speaker’s FOA is not directed at any other participant (1/3 of the data) and all utterances during which the speaker’s FOA is directed at least once at another participant (2/3 of the data).

First Experiment For the first experiment we trained a Bayesian Network adapted from Jovanovič (2007) on a mix of utterances from all contexts, and tested its performance on utterances from the three different contexts: (1) all data, (2) all data in the context ‘at least some person in speaker’s FOA’ and (3) all data in the context ‘no person in speaker’s FOA during utterance’. As was to be expected, the performance in the second context showed a clear gain compared to the first context, and the performance in the third context was clearly worse. The performance differences, for different train/test splits, tend to be about five percent.

Second Experiment Because the second context showed such a better performance, we ran a second experiment where we trained the network on only data from the second context, to see if we could improve the performance in that context even more. In different train/test splits this gave us another small performance increase.

Conclusions for Contextual Addressing The performance increases can mostly be attributed to the distinction between different individual addressees for I-addressed utterances. Precision and recall for the G-addressed utterances does not change so much for the different contexts. This result is reminiscent of the fact that when the annotators agreed on an utterance being I-addressed they typically also agreed on the particular individual being addressed.

These results are particularly interesting in the light of the high accuracy with which FOA was annotated. If this accuracy points at the possibility to also achieve a high automatic recognition rate for FOA we can exploit these results in a practical application context by defining a addressee detection module which only assigns an addressee to an utterance in the second FOA context (FOA at some participants), and in all other cases labels an utterance as ‘addressee cannot be determined’. Such a detection module achieves a much higher precision than a module that tries to assign an addressee label regardless; of course this happens at the cost of recall.

2.2 Interannotator Training and Testing

Classifiers behave as they are trained. When two annotators differ in the way they annotate, i.e. have different “mental conceptions” of the phenomenon being annotated, we can expect that a classifier trained on the data annotated by one annotator behaves different from a classifier trained on the other annotator’s data. As Rienks describes, this property allows us to use all data in the corpus, instead of just the multiply annotated part of it, for analyzing differences between annotators (Rienks, 2007, page 105). We can expect that a classifier A trained on data annotated by A will perform better when tested on data annotated by A, than when tested on data annotated by B. In other words, classifier A is geared towards modelling the ‘mental conception’ of annotator A. In this section we will try to find out whether it is possible to explicitly tease apart the overlap and the differences in the mental conceptions of the annotators as mirrored in the behavior of classifiers, on a subjective annotation task. Suppose that we build a Voting Classifier, based on the votes of a number of classifiers each trained on a different annotator’s data. The Voting Classifier only makes a decision when all voters agree on the class label. How good will the Voting Classifier perform? Is there any relation between the (dis)agreement of the voters, and the (dis)agreement of the annotators? Will the resulting Voting Classifier in some way embody the overlap between the ‘mental conceptions’ of the different annotators?

As an illustration and a test case for such a Voting Classifier, we consider the human annotations and automatic classification of a particular type of utterances in the AMI corpus, the *“Yeah-utterances”*, utterances that start with the word “yeah”.

class	train-tot	test-tot	DH-train/test	S9-train/test	VK-train/test
bc	3043	1347	1393/747	670/241	980/359
as	3724	1859	1536/1104	689/189	1499/566
in	782	377	340/229	207/60	235/88
ot	1289	596	316/209	187/38	786/349

Table 1: Sizes of train and test data sets used and the distribution of class labels over these data sets for the different annotators.

The Data Response tokens like "yeah", "okay", "right" and "no" have the interest of linguists because they may give a clue about the stance that the listener takes towards what is said by the speaker (Gardner, 2004). Jefferson described the difference between "yeah" and other backchannels in terms of speaker recipiency, the willingness of the speaker to take the floor (Jefferson, 1984). Yeah utterances make up a substantial part of the dialogue acts in the AMI meeting conversations (about eight percent). "Yeah" is the most ambiguous utterance that occurs in discussion segments in AMI meetings. In order to get information about the stance that participants take with respect towards the issue discussed it is important to be able to tell utterances of "Yeah" as a mere backchannel, from Yeah utterances that express agreement with the opinion of the speaker (see the work of Heylen and Op den Akker (2007)).

The class variables for dialogue act types of Yeah utterances that are distinguished are: Assess (as), Backchannel (bc), Inform (in), and Other (ot). Table 1 gives a distribution of the labels in our train and test data sets. Note that for each annotator, a disjunct train and test set have been defined. The inter-annotator agreement on the Yeah utterances is low. The pairwise alpha values for meeting IS1003d, which was annotated by all three annotators, are (in brackets the number of agreed DA segments that start with "Yeah"): alpha(VK,DH) = 0.36 (111), alpha (VK,S9) = 0.36 (132), alpha(DH,S9) = 0.45 (160).

Testing for Systematic Differences When one suspects the annotations to have originated from different mental conceptions of annotators, the first step is to test whether these differences are systematic. Table 2 presents the intra and inter annotator classification accuracy. There is a clear performance drop between using the test data from the same annotator from which the training data was taken and using the test data of other annotators or the mixed test data of all annotators. This suggest that some of the disagreements in the annotation stem from systematic differences in the mental conceptions of the annotators.

	TEST			
TRAIN	DH	S9	VK	Mixed
DH	**69**	64	52	63
S9	59	**68**	48	57
VK	63	57	**66**	63

Table 2: Performance of classifiers (in terms of accuracy values – i.e. percentage correct predictions) trained and tested on various data sets. Results were obtained with a decision tree classifier, J48 in the Weka toolkit.

Building the Voting Classifier Given the three classifiers DH, S9 and VK, each trained on the train data taken from one single annotator, we have build a Voting Classifier that outputs a class label when all three 'voters' (the classifiers DH, S9 and VK) give the same label, and the label 'unknown' otherwise. As was to be expected, the *accuracy* for this Voting Classifier is much lower than the accuracy of each of the single voters and than the accuracy of a classifier trained on a mix of data from all annotators (see Table 3), due to the many times the Voting Classifier assigns the label 'unknown' which is not present in the test data and is always false. The precision of the Voting Classifier however is higher than that of any of the other classifiers, for each of the classes (see Table 4).

Conclusions for the Voting Classifier For the data that we used in this experiment, building a Voting Classifier as described above gave us a high precision classifier. Based on our starting point, this would relate to the classifier in some way embodying the overlap in the mental conceptions of each of the annotators. If that were true, the cases in which the Voting Classifier returns an unanimous vote would be mostly those cases in which the different annotators would also have agreed.

TRAIN	Accuracy
train_MIX(8838)	67
DH(3585)	63
S9(1753)	57
VK(3500)	63
VotingClassifier(8838)	43

Table 3: Performance of the MaxEnt classifiers (in terms of accuracy values – i.e. percentage correct predictions) tested on the whole test set, a mix of three annotators data (4179 "Yeah" utterances). The first column between brackets the size of the train sets.

	Classifier				
Class	Voting	DH	S9	VK	train_MIX
BC	71	65	63	71	69
AS	73	62	64	61	66
IN	60	58	34	52	50
OT	86	59	32	57	80

Table 4: Precision values per class label for the classifiers.

This can be tested quite simply using multiply annotated data. Note that not *all* data needs to be annotated by more annotators: just enough to test this hypothesis. Otherwise, it will suffice to have enough data for each single annotator, be it overlapping or not. This is especially advantageous when the corpus is really large, such as the 100h AMI corpus. Another way to test the hypothesis that the voting behavior relates to intersubjectivity is to look at the type and context of the agreements between annotators, found in the reliability analysis, and see if that relates to the type and context of the cases where the Voting Classifier renders an unanimous judgement. That would be strong circumstantial evidence in support of the hypothesis.

Note that the gain in precision is obtained at the cost of recall, because the Voting Classifier approach explicitly restricts judgements to the cases where annotators would have agreed and, presumably, therefore to the cases in which users of the data are able to agree to the judgements as well. It is possible that you 'lose' a class label in the classifier by having a high precision but a recall of less than five percent, which in our example happened for the 'other' class.

3 The Classifier as Subjective Entity vs the Classifier as Embodiment of Consensus Objectivity

Many annotation tasks are subjective to a larger degree. When this is simply taken as a given, and the systematic disagreements resulting from the different mental conceptions of the annotators are not taken into account while training a machine classifier on the resulting data, there is no simple reason to assume that the resulting classifier is any less subjective in the judgements it makes. Without additional analyses one cannot suppose the classifier did not pick up idiosyncrasies from the annotators. We have seen that machine classifiers can indeed considered to be subjective in their judgements, a property they have inherited from the annotations they have been trained on. A judgement made by such a classifier should be approached in a similar manner as a judgement made by another person[3]. We will call the resulting classifier therefore a *'subjective entity'* classifier.

A careful analysis of the interannotator agreements and disagreements might make it possible to build classifiers that partly embody the intersubjective overlap between the mental conceptions of the annotators. Because the classifier only tries to give a judgement in situations where one can expect annotators or users to agree, one can approach the judgements made by the classifier as a "common sense" of judgements that people can agree on, despite the subjective quality of the annotation task. We will call the resulting classifier a *'consensus objective'* classifier.

4 Discussion

In the Introduction we distinguished several uses of data annotation using human annotators. The analyses and research in this paper mainly concerns the use of annotated data for the training and development of automatic machine classifiers. Ideally the annotation schema and the class labels that are distinguished reflect the use that is made of the output of the machine classifiers in some particular application in which the classifier operates as a module. Imagine for example a system that detects when meeting participants are too dominant and signals the chairman of the meet-

[3] On a side note, letting the machine classifiers judgments be presented through an embodied conversational agent can be a way to present this human-like subjectivity for the user (Reidsma et al., 2007).

ing to prevent some participants being dissatisfied with the decision making processes. Or, a classifier for addressee detection that signals remote participants that they are addressed by the speaker. The way that users of the system interpret the signals output by the classifier should correspond to the meanings that were used by the annotators and that were implemented in the classifier.

When there is a lot of disagreement in the annotations this should be taken into account for machine learning if one does not want to obtain a ‘subjective entity’ classifier, the judgements of which the user will often disagree with. In Section 2 we presented two ways to exploit such data for building machine classifiers. Here we elaborate a bit on a difference between the two cases relating to the different *causes* of the inter-annotator disagreement.

For the addressing annotations, the annotators sometimes had problems with choosing between G-addressed and I-addressed. The *participants* in the conversation usually did not seem to have any problem with that. There are only a few instances in the data where the participants explicitly requested clarification. It is reasonable to expect that in cases where it really matters – for the conversational partners – who is being addressed, outside observers will not have a problem to identify this. Thus, in those cases where the annotators had problems to decide upon the type of addressing there maybe was no reason for the participants in the conversation to make that clear because it simply was not an issue. The annotators were then tripped by the fact that they were *forced* by the annotation guidelines to choose one addressee label.

In the dialogue act classification task something additional is going on. Here we see that annotators also have problems because many utterances themselves are ambiguous or poly-interpretable. Some annotator may prefer to call this act an assess where an other prefers to call it an inform, and both may have good reason to back up their choice. A similar situation occurs in the case of the classification of Yeah utterances. The disagreements then seem to be caused more explicitly by differing judgements of a conversational situation.

5 Conclusions

We have argued that dis-agreements between different observers of ‘subjective content’ is unavoidable and an intrinsic quality of the interpretation and classification process of such type of content. Any subdivision of these type of phenomena into a predefined set of disjunct classes suffers from being arbitrary. There are always cases that can belong to this but also to that class. Analysis of annotations of the same data by different annotators may reveal that there are differences in the decisions they make, such as some personal preference for one class over another.

Instead of throwing away the data as not being valuable at all for machine learning purposes, we have shown two ways to exploit such data, both leading to high precision / low recall classifiers that in some cases refuse to give a judgement. The first way was based on the identification of subsets of the data that show higher inter-annotator agreement. When the events in these subsets can be identified computationally the way is open to use classifiers trained on these subsets. We have illustrated this with several subsets of addressing events in the AMI meeting corpus and we have shown that this leads to an improvement in the accuracy of the classifiers. Precision is raised in case the classifier refrains from making a decision in those situation that fall outside the subsets. The second way is to train a number of classifiers, one for each of the annotators data part of the corpus, and build a Voting Classifier that only makes a decision in case all classifiers agree on the class label. This approach was illustrated by the problem of classification of the dialogue act type of Yeah-utterances in the AMI corpus. The results show that the approach indeed leads to the expected improvement in precision, at the cost of a lower recall, because of the cases in which the classifier doesn’t make a decision.

Acknowledgements

The authors are in debt to many people for many fruitful discussions, most prominently Jean Carletta, Ron Artstein, Arthur van Bunningen, Henning Rode and Dirk Heylen. This work is supported by the European IST Programme Project FP6-033812 (AMIDA, publication 136). This article only reflects the authors’ views and funding agencies are not liable for any use that may be made of the information contained herein.

References

Ba, S. O. and J.-M. Odobez. 2006. A study on visual focus of attention recognition from head pose in a meeting room. In Renals, S. and S. Bengio, editors, *Proc. of the MLMI 2006*, volume 4299 of *Lecture Notes in Computer Science*, pages 75–87. Springer.

Bakeman, R. and J. M. Gottman. 1986. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.

Carletta, J. C. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Craggs, R. and M. McGee Wood. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3):289–296.

Galley, M., K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proc. of the 42nd Meeting of the ACL*, pages 669–676. ACL.

Gardner, R. 2004. Acknowledging strong ties between utterances in talk: Connections through right as a response token. In *Proceedings of the 2004 Conference of the Australian Linguistic Society*, pages 1–12.

Goffman, E. 1981. Footing. In *Forms of Talk*, pages 124–159. Philadelphia: University of Pennsylvania Press.

Heylen, D. and H. op den Akker. 2007. Computing backchannel distributions in multi-party conversations. In Cassell, J. and D. Heylen, editors, *Proc. of the ACL Workshop on Embodied Language Processing, Prague*, pages 17–24. ACL.

Jefferson, G. 1984. Notes on a systematic deployment of the acknowledgement tokens ‘yeah’ and ‘mm hm’. *Papers in Linguistics*, 17:197–206.

Jovanović, N., H. op den Akker, and A. Nijholt. 2006. A corpus for studying addressing behaviour in multiparty dialogues. *Language Resources and Evaluation*, 40(1):5–23.

Jovanović, N. 2007. *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*. Phd thesis, University of Twente.

Krippendorff, K. 1980. *Content Analysis: An Introduction to its Methodology*, volume 5 of *The Sage CommText Series*. Sage Publications, Beverly Hills, London.

Passonneau, R. J., T. Yano, T. Lippincott, and J. Klavans. 2008. Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. In *Proc. of the LREC 2008*.

Poesio, M. and R. Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. ACL.

Potter, J. W. and D. Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of applied communication research*, 27(3):258–284.

Reidsma, D. and J. C. Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3).

Reidsma, D., Z. M. Ruttkay, and A. Nijholt, 2007. *Challenges for Virtual Humans in Human Computing*, chapter 16, pages 316–338. Number 4451 in LNAI: State of the Art Surveys. Springer Verlag, Berlin/Heidelberg.

Reidsma, D., D. Heylen, and H. op den Akker. 2008. On the contextual analysis of agreement scores. In *Proc. of the LREC Workshop on Multimodal Corpora*.

Rienks, R. J. 2007. *Meetings in Smart Environments: Implications of progressing technology*. Phd thesis, SIKS Graduate School / University of Twente, Enschede, NL.

Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann. 2005. “of all things the measure is man” automatic classification of emotion and intra labeler consistency. In *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing*.

Vieira, R. 2002. How to evaluate systems against human judgment on the presense of disagreement? In *Proc. workshop on joint evaluation of computational processing of Portugese at PorTAL 2002*.

Voorhees, E. M. and D. Harman. 1997. Overview of the trec-5. In *Proc. of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28. NIST.

Voorhees, E. M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716.

Wiebe, J. M., R. F. Bruce, and T. P. O’Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proc. of the 37th Annual Meeting of the ACL*, pages 246–253. ACL.

Wilson, T. 2008. Annotating subjective content in meetings. In *Proc. of the Language Resources and Evaluation Conference (LREC-2008)*.

Human judgment as a parameter in evaluation campaigns

Jean-Baptiste Berthelin and **Cyril Grouin** and **Martine Hurault-Plantet** and **Patrick Paroubek**
LIMSI-CNRS
BP 133
F-91403 Orsay Cedex
firstname.lastname@limsi.fr

Abstract

The relevance of human judgment in an evaluation campaign is illustrated here through the DEFT text mining campaigns.

In a first step, testing a topic for a campaign among a limited number of human evaluators informs us about the feasibility of a task. This information comes from the results obtained by the judges, as well as from their personal impressions after passing the test.

In a second step, results from individual judges, as well as their pairwise matching, are used in order to adjust the task (choice of a marking scale for DEFT'07 and selection of topical categories for DEFT'08).

Finally, the mutual comparison of competitors' results, at the end of the evaluation campaign, confirms the choices we made at its starting point, and provides means to redefine the task when we shall launch a future campaign based on the same topic.

1 Introduction

For the past four years, the DEFT[1] (*Défi Fouille de Texte*) campaigns have been aiming to evaluate methods and software developed by several research teams in French text mining, on a variety of topics.

The different editions concerned, in this order, the identification of speakers in political speeches (2005), the topical segmentation of political, scientific and juridical corpora (2006), the automatic affectation of opinion values to texts developing an argumented judgment (2007), and the identification of the genre and topic of a document (2008).

Human judgment was used during the preparation of the last two campaigns, to assess the difficulty of the task, and to see which parameters could be modified. To do this, before the participants start competing via their software, we put human judges in front of versions of the task with various sets of parameters. This allows us to adjust the definition of the task according to which difficulties were encountered, and how judges agree together. These human judges are in small number, and belong to our team. However, results of the campaign are automatically evaluated with reference to results attached to the corpus from the start. This is because the evaluation of a campaign's results by human judges is expensive. For instance, TREC[2] international evaluation campaigns are supported by the NIST institute and funded by state agencies. In Europe, on the same domains, the CLEF[3] campaigns are funded by the European Commission, and in France, evaluation campaigns are also funded by projects, such as Technolangue[4]. DEFT campaigns, however, are conducted with small budgets. That means for us to have selected corpora that contain the desired results. For instance, in a campaign for topical categorization, we must start with a topically tagged corpus. By so doing, we also can, at the end of a campaign, compare results from human judges with results from competitors, using an identical

[1] See http://deft.limsi.fr/ for a presentation in French.

[2] http://trec.nist.gov

[3] http://www.clef-campaign.org

[4] http://www.technolangue.net

Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, pages 17–23
Manchester, August 2008

common reference.

In this paper, we describe experiments we performed with human judgments when preparing DEFT campaigns. We survey the various steps in the preparation of the last two campaigns, and we go through the detail of how human evaluation, performed during these steps, led us to the parametrization of these two campaigns. We also present a comparative analysis of results found by human judges and results submitted by competitors in the challenge. We conclude about the relevance of the human evaluation of a task, prior to evaluating software dedicated to this task.

2 Parametrization of the campaign

We were competitors in the 2005 and 2006 editions, and became organisators for the 2007 and 2008 campaigns. For both challenges that we organized, we went through the classical steps of the evaluation paradigm (Adda et al., 1999), to which we systematically added a step of human test of the task, in order to adjust those parameters that could be modified. The steps, therefore, are following:

1. thinking about potential topics;
2. choice of a task and collection of corpora;
3. choice of measurements;
4. test of the task by human judges on an extract of the corpus in order to precisely define its parameters;
5. launching the task, recruiting participants;
6. testing period;
7. adjudication: possibility of complaints about the results;
8. workshop that closes the campaign.

Whenever human judges have to evaluate the results of participants in a campaign, the main problems are about correctly defining the judging criteria to be applied by judges, and that judges be in sufficient number to vote on judging each document. Hovy et al. (2002) describe work toward formalization of software evaluation methodology in NLP, developed in the EAGLES[5] and ISLE[6] projects. For cost-efficiency reasons, automatic evaluation is relevant, and its results have sometimes been compared to results from human judges. For instance, Eck and Hori (2005) compare results of evaluation measurements used in automatic translation with human judgments on the same corpus. In (Burstein and Wolska, 2003), the authors describe an experiment in the evaluation of writing style and find a better agreement between the automatic evaluation system and one human judge, than between two human judges.

Returning to the DEFT campaign, once the task is chosen, the corpora are collected, and evaluation measurements are defined, there can remain some necessity of adjusting parameters, according to the expected difficulty of the task. This could be, for instance, the level of granularity in a task of topical segmentation, or which categories should be relevant in a task of categorization. To get this adjusting done, we submit the task to human judges.

In 2007, the challenge was about the automatic affectation of opinion values to texts developing an argumented judgment (Grouin et al., 2007). We collected opinion texts already tagged by an opinion value, such as film reviews that, in addition to a text giving the judgment of the critic on the film, also feature a mark in the shape of a variable number of stars. The adjustable parameter of the task, therefore, is the scale of opinion values. The task will be more or less difficult, according to the range of this scale.

The 2008 campaign was about classifying a set of documents by genre and topic (Hurault-Plantet et al., 2008). The choice of genres and topics is a crucial one. Some pairs of topics or genres are more difficult to separate than other ones. We also had to find different genres sharing a set of topical categories, while corpora in French are not so very abundant. So we selected two genres, encyclopedia and daily newspaper, and about ten general topical categories. The parameter we had to adjust was the set of categories to be matched against each other.

3 Assessing the difficulty of a task

3.1 Calibration of an opinion value scale

In 2007, the challenge was about the automatic affectation of opinion values to texts developing an argumented judgment. In view of that, we collected four corpora that covered various domains:

[5] http://www.ilc.cnr.it/EAGLES96/home.html

[6] http://www.ilc.cnr.it/EAGLES96/isle/

reviews of films and books, of video games and of scientific papers, as well as parliamentary debates about a draft law.

Each corpus had the interesting feature of combining a mark or opinion with a descriptive text, as the mark was used to sum up the judment in the argumentative part of this text. Due to the diversity of sources, we found as many marking scales as involved copora:

- 2 values for parliamentary debates[7] (the representative who took part in the debate was either in favour or in disfavour of the draft law) ;
- 4 values for scientific paper reviews (*accepted as it stands – accepted with minor changes – accepted with major changes and second overall reviewing –rejected*), based on a set of criteria including interestingness, relevance and originality of the paper's content ;
- 5 values for film and book reviews[8] (a mark between 0 and 4, from bad to excellent) ;
- 20 values for video game reviews[9] (a global mark calculated from a set of advices about various aspects of the game: graphics, playability, life span, sound track and scenario).

In order to, first, assess the feasibility of the task, and to, secondly, define the scale of values to be used in the evaluation campaign, we submitted human judges to several tests (Paek, 2001): they were instructed to assign a mark on two kinds of scale, a wide one with the original values, and a restricted one with 2 or 3 values, depending on the corpus it was applying to. The results from various judges were evaluated in terms of precision and recall, and matched to each other by way of the Kappa coefficient (Carletta, 1996) (Cohen, 1960).

We present hereunder the values of the κ coefficient between pairs of human judges, and with the reference, on the video game corpus. The wide scale (Table 1) uses the original values (marks between 0 and 20), while the restricted scale (Table 2) relies upon 3 values with following definitions: class 0 for original marks between 0 and 10, class 1 for marks between 11 and 14, and class 2 for marks between 15 and 20.

Judge	**Ref.**	**1**	**2**	**3**
Ref.		0.17	0.12	0.07
1	0.17		0.03	0.05
2	0.12	0.03		0.07
3	0.07	0.05	0.07	

Table 1: Video game corpus: wide scale, marks from 0 to 20.

Judge	**Ref.**	**1**	**2**	**3**
Ref.		0.74	0.79	0.69
1	0.74		0.74	0.54
2	0.79	0.74		0.69
3	0.69	0.54	0.69	

Table 2: Video game corpus: restricted scale, marks from 0 to 2.

Table 1 and 2 show that agreement between judges varies widely when marking scales are modified. Table 1 shows that there is an insufficient agreement among judges on the wide scale, with κ coefficients lower than 0.20, while the agreement between these same judges can be considered as good on the restricted scale, with κ coefficients between 0.54 and 0.79 (Table 2), the median being at 0.74.

In order to confirm the validity of the change in scales, we used the κ to test how each judge agreed with himself, between his two sets of results (Table 3). Therefore, we compared judgments made by each judge using the initial value scale and converted towards the restricted scale, with judgments made by the same judge directly using the restricted value scale. This measurement shows the degree of correspondence between both scales for each judge. Among the three judges who took part in the test, the first and third one agree well with themselves, while for the second one, the agreement is only moderate.

Judge	**1**	**2**	**3**
1	0.74		
2		0.46	
3			0.70

Table 3: Video game corpus: agreement of each judge with himself when scales change.

We did the same for a second corpus, of film reviews. The test involved five judges, and the scale

[7] `http://www.assemblee-nationale.fr/12/debats/`

[8] `http://www.avoir-alire.com`

[9] `http://www.jeuxvideo.com/etajvbis.htm`

change was smaller, since it was from five values to three, and not from twenty to three. For this scale change, we merged the two lowest values (0 and 1) into one (0), and the two highest ones (3 and 4) into one (2), and the middle value in the wide scale (2) remained the intermediate one in the restricted scale (1). This scale change was the most relevant one, since, with 29.7% of the documents, the class of the middle mark (2) accounted for almost one third of the corpus. However, the two other groups of documents are less well balanced. Indeed, the lowest mark concerns less documents than the highest one: 4.6% and 10.3% respectively for the initial marks 0 and 1, while one finds 39.8% and 15.6% of documents for the marks 3 and 4. Grouping the documents in only two classes, by joining the middle class with the two lowest ones, would have yielded a better balance between classes, with 44.6% of documents for the lower mark and 55.4% for the higher one, but that would have been less meaningful.

Results from human judges are shown in the Tables 4 and 5 for both scales.

Judge	**Ref.**	**1**	**2**	**3**	**4**	**5**
Ref.		0.10	0.29	0.39	0.46	0.47
1	0.10		0.37	0.49	0.48	0.35
2	0.29	0.37		0.36	0.30	0.43
3	0.39	0.49	0.36		0.49	0.54
4	0.46	0.48	0.30	0.49		0.60
5	0.47	0.35	0.43	0.54	0.60	

Table 4: Film review corpus: wide scale, marks from 0 to 4

Judge	**Ref.**	**1**	**2**	**3**	**4**	**5**
Ref.		0.27	0.62	0.53	0.56	0.67
1	0.27		0.45	0.43	0.57	0.37
2	0.62	0.45		0.73	0.48	0.54
3	0.53	0.43	0.73		0.62	0.62
4	0.56	0.57	0.48	0.62		0.76
5	0.67	0.37	0.54	0.62	0.76	

Table 5: Film review corpus: restricted scale, marks from 0 to 2.

Agreements between human judges ranked from bad to moderate for the wide scale (the five original values in this corpus), while these agreements rank from insufficient to good in the case of the restricted scale with three values. We can see that differences induced by the scale change are much less important than with the video game corpus. This agrees well with the scales being much closer to each other.

By first performing a hand-made evaluation, and secondly, matching between themselves the results from the judges, we found a way to assess with greater precision the difficulty of the evaluation task we were about to launch. Concerning the first two review corpora (films and books, video games), we attached values good, average and bad to the three selected classes. The scale for scientific paper reviews was also restricted to three classes for which following values were selected: paper accepted as it stands or with minor edits, paper accepted after major edits, paper rejected. Finally, since its original scale had only two values, the corpus of parliamentary debates underwent no change of scale.

3.2 Choice of a topical category set

In order to determine which topical categories should be recognized in the 2008 task of classifying documents by genre and topic, we performed a manual evaluation of a sample of the corpus with 4 human judges. The sample included 30 Le Monde papers for the journalistic genre, and 30 Wikipedia entries for the encyclopedic genre. Only the title and body of each article was kept in the sample, and the tables were deleted. All marks of inclusion in either corpus were also deleted (references to Le Monde and Wikipedia tags).

The test ran this way: each article was put in a separate file, and the evaluators had to identify the genre and the topical category under which it was published. All articles were included in one set, which means evaluators had to choose, between all categories and genres, which ones to match with each document. This test was made with a first selection of 8 categories, shared by both genres, listed in Table 6.

Table 7 shows that results from human judges in terms of precision and recall were excellent on the identification of genre (F-scores between 0.94 and 1.00) and quite good on the identification of categories (F-scores between 0.66 and 0.82).

We also proceeded to the pairwise matching of results from human judges via the κ coefficient. Results show an excellent agreement of judges among themselves and with the reference for genre identification (Table 8). The agreement is mod-

Le Monde	**Wikipedia**
Notebook	*People*
Economy	*Economy*
France	*French Politics*
International	*International Politics,* minus category *French Politic*
Science	*Science*
Society	*Society,* minus subcategories *Politics, People,* *Sport, Media*
Sport	*Sport*
Television	*Television*

Table 6: Correspondence between categories from Le Monde and Wikipedia for the 8 categories in the test.

Judge	**1**	**2**	**3**	**4**
Genres	1.00	0.98	0.97	0.94
Categories	0.79	0.77	0.82	0.66

Table 7: F-scores obtained by human judges on the identification of genre and categories.

erate to good for categoy identification (Table 9). These good results led us to keep the corpora as they stood, since they appeared to constitute a good reference for the defined task. However, we made an exception for category *Notebook* (biographies of celebrities) which we discarded for two reasons. First, it is more of a genre, namely, ”biography”, rather than a topical category. Secondly, we found it rather difficult to assign a single category to articles which could belong in two different ones, as would be the case for the biography of a sportsman, which would fall under both categories *Notebook* et *Sport*.

Judge	**Réf.**	**1**	**2**	**3**	**4**
Réf.		1.00	0.97	0.93	0.87
1	1.00		0.97	0.93	0.87
2	0.97	0.97		0.90	0.83
3	0.93	0.93	0.90		0.87
4	0.87	0.87	0.83	0.87	

Table 8: κ coefficient between human judges and the reference: Identification of genre.

Judge	**Réf.**	**1**	**2**	**3**	**4**
Réf.		0.56	0.52	0.60	0.39
1	0.56		0.69	0.75	0.55
2	0.52	0.69		0.71	0.61
3	0.60	0.75	0.71		0.52
4	0.39	0.55	0.61	0.52	

Table 9: κ coefficient between human judges and the reference: Identification of categories.

Our task of genre and topic classification included two subtasks, one being genre and topic recognition for a first set of categories, the other one being only topic recognition for a second set of categories. Therefore, the corpus had to be divided in two parts. In order to find which categories had to go into which subcorpus, we decided to estimate, for each category, the difficulty of recognizing it. To do so, we calculated the precision and recall of each evaluator for each category. This measurement was obtained via a second evaluation of human judges, with a wider set of categories (by adding categories *Art* and *Literature*).

The ordering of categories by decreasing precision is following: *Sport* (1.00), *International* (0.80), *France* (0.76), *Literature* (0.76), *Art* (0.74), *Television* (0.71), *Economy* (0.58), *Science* (0.33), *Society* (0.26). This means no document in the *Sport* category was misclassified, and, contrariwise, categories *Science* and *Society* were the most problematic ones.

The ordering by decreasing recall is slightly different: *International* (0.87), *Economy* (0.80), *Sport* (0.75), *France* (0.70), *Art* (0.62), *Literature* (0.49), *Television* (0.46), *Society* (0.42), *Science* (0.33). Hence, articles in the *International* category were best identified. This ordering also confirms the difficulty felt by human judges concerning the categories *Society* and *Science*.

We decided to distribute the categories for each subtask according to a balance between easy and diffucult ones in terms of human evaluation:

- *Art, Economy, Sport, Television* for the subtask with both genre and category recognition;

- *France, International, Literature, Science, Society* for the subtask with only category recognition. For this second subset, we put together three categories which are topically close (*France, International* and *Society*).

4 Human judgments and software

4.1 Confirming the difficulty of a task

The 2007 edition of DEFT highlighted two main phenomena concerning the corpora involved in the task.

First, each corpus yielded a different level of difficulty, and this gradation of difficulty among corpora appeared both for human evaluators and competitors in the challenge (Paroubek et al., 2007).

	Judges	**Competitors**
Debates	0.77/1.00	0.54/0.72
Game reviews	0.73/0.90	0.46/0.78
Film reviews	0.52/0.79	0.38/0.60
Paper reviews	0.41/0.58	0.40/0.57

Table 10: Minimal and maximal strict F-scores between human evaluators and competitors in the challenge, 2007 edition.

During human tests, judges mentioned the great facility of finding about opinions expressed in the corpus of parliamentary debate. Next came corpora of video game reviews, and then of film and book reviews, whose difficulty was considered average, and last, the corpus of scientific paper reviews, which the judges perceived as particularly difficult. This gradation of difficulty among corpora was also found among competitors, following the same ordering of three levels of difficulty.

Secondly, the difficulties met by human evaluators are also found in the case of competitors. Upon finishing human tests, judges felt difficulties in evaluating the corpus of scientific paper reviews, yielding poor results. Now, the results of competitors on the same corpus are quite as poor, occupying exactly the same value interval as for human judges. Most competitors, by the way, obtained their worst results on this corpus.

The alikeness of results between judges and competitors reflects the complexity of the corpus: when preparing the campaign, we observed that reviews were quite short. Therefore, assigning a value had to rely upon a small amount of data. From that, we can derive a minimal size for documents to be used in this kind of evaluation. Moreover, a paper review can be seen as an aid for the author, to be expressed as positively as possible, even if it is also addressed to the Program Committee which has to accept or reject the paper. Therefore, the mark could prove more negative than the text of the review.

The case of comments about videogames is a different one. Indeed, giving a global mark on a scale of 20 is a difficult task. Therefore, this mark comes most often from a sum of smaller marks which rate either the whole document according to various criteria, or parts of this document. In our corpus, each reviewer rates the game according to several criteria, namely, graphics, playability, life span, sound track and scenario, from which a rather long text is produced, making the judgment an easier task to perform. However, the global mark differs from the sum of the smaller ones from various criteria, hence the difficulty for human judges to reckon this global mark on a scale of 20.

4.2 Confirmation of the expected success of competitors

Contrary to the 2007 edition, in which competitors obtained results that confirmed those of human judges, the 2008 edition gave them the opportunity to reach a higher level than human evaluators.

While genre identification yielded no special problem, either for human evaluators or for competitors, and the results obtained by both groups are similar, competitors reached better results than human judges in topical categorization.

Concerning genre identification, strict F-scores are situated between 0.94 and 1.00 for human judges, and between 0.95 and 0.98 for the best runs of competitors (each competitor was allowed to submit up to three collections of results, only the best one being used for the final ranking). As for topical categorization, strict F-scores go from 0.66 to 0.82 for human evaluators, and from 0.84 to 0.89 for best runs from competitors.

The equivalence of results on genre identification between judges and competitors can be explained by the fact that it was a simple, binary choice (the newspaper Le Monde vs. Wikipedia).

Contrariwise, competitors obtained better results in topical categorization, since machines have a stronger abstraction capacity than humans in presence of the 9 topical categories we defined (*Art, Economy, France, International, Literature, Science, Society, Sport* and *Television*). However, conditions were not quite similar, since human judges had to pick a category among eight, and not, like the automatic systems, a category within two subsets of four and five categories. Indeed, we dispatched the categories into two sets, by balancing categories that are easy or difficult for human evaluators. For the second set of categories, we carefully put together three semantically close ones, (*France, International* and *Society,* all three of them being about political and societal contents), to make the task more difficult. Although the second set of categories seems more complicated for human judges, half of the competitors obtained better results in topical categorization of the second set than of the first one.

5 Conclusion

The relevance of human judgment in an evaluation campaign is present from the beginning to the end of a campaign.

In a first step, testing a topic for a campaign among a limited number of human evaluators allows us to check the feasibility of a task. This checking relies both on the results obtained by judges (recall, precision, F-scores) and on their personal impressions after passing the test.

In a second step, the study of both the results obtained by the judges, and their pairwise matching involving such a comparator as the κ coefficient allows us to adjust the task (choice of a marking scale for DEFT'07 and selection of topical categories for DEFT'08).

Finally, the mutual comparison of competitors' results, at the end of the evaluation campaign, allows us to validate the choices we made at its starting point, and even to reposition the task when we shall launch a future campaign based on the same topic.

References

Adda, Gilles, Joseph Mariani, Patrick Paroubek, Martin Rajman, and Josette Lecomte. 1999. L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, 2(2):119–129, juin.

Burstein, Jill and Magdalena Wolska. 2003. Toward evaluation of writing style: Finding overly repetitive word use in student essays. In *10th Conference of the European Chapter of the Association for Computational Linguistics, EACL'03*, pages 35–42, Budapest, Hungary, april.

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

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

Eck, Matthias and Chiori Hori. 2005. Overview of the iwslt 2005 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 5–14, Pittsburg, PA.

Grouin, Cyril, Jean-Baptiste Berthelin, Sarra El Ayari, Thomas Heitz, Martine Hurault-Plantet, Michèle Jardino, Zohra Khalis, and Michel Lastes. 2007. Présentation de DEFT'07 (DÉfi Fouille de Textes). In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 1–8, Grenoble. Association Française d'Intelligence Artificielle.

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*.

Hurault-Plantet, Martine, Jean-Baptiste Berthelin, Sarra El Ayari, Cyril Grouin, Patrick Paroubek, and Sylvain Loiseau. 2008. Résultats de l'édition 2008 du DÉfi Fouille de Textes. In *Actes TALN'08*, Avignon. Association pour le Traitement Automatique des Langues.

Paek, Tim. 2001. Empirical Methods for Evaluating Dialog Systems. In *Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*, pages 3–10.

Paroubek, Patrick, Jean-Baptiste Berthelin, Sarra El Ayari, Cyril Grouin, Thomas Heitz, Martine Hurault-Plantet, Michèle Jardino, Zohra Khalis, and Michel Lastes. 2007. Résultats de l'édition 2007 du DÉfi Fouille de Textes. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, pages 9–17, Grenoble. Association Française d'Intelligence Artificielle.

Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection

Joel R. Tetreault
Educational Testing Service
660 Rosedale Road
Princeton, NJ, USA
JTetreault@ets.org

Martin Chodorow
Hunter College of CUNY
695 Park Avenue
New York, NY, USA
martin.chodorow@hunter.cuny.edu

Abstract

Evaluation and annotation are two of the greatest challenges in developing NLP instructional or diagnostic tools to mark grammar and usage errors in the writing of non-native speakers. Past approaches have commonly used only one rater to annotate a corpus of learner errors to compare to system output. In this paper, we show how using only one rater can skew system evaluation and then we present a sampling approach that makes it possible to evaluate a system more efficiently.

1 Introduction

In this paper, we present a series of experiments that explore the reliability of human judgments in rating preposition usage. While one tends to think of annotator disagreements about discourse and semantics as being quite common, our studies show that judgments of preposition usage, which is largely lexically driven, can be just as contentious. As a result, this unreliability poses a serious issue for the development and evaluation of NLP tools in the task of automatically detecting preposition usage errors in the writing of non-native speakers of English.

To date, single human annotation has typically been the gold standard for grammatical error detection, such as in the work of (Izumi et al., 2004), (Han et al., 2006), (Nagata et al., 2006), (Gamon et al., 2008)[1]. Although there are several learner corpora annotated for preposition and determiner errors (such as the Cambridge Learners Corpus[2] and the Chinese Learner English Corpus[3]), it is unclear which portions of these, if any, were doubly annotated. This previous work has side-stepped the issue of annotator reliability, which we address here through the following three contributions:

- **Judgments of Native Usage** To motivate our work in non-native usage, we first illustrate the difficulty of preposition selection with two experiments: a cloze test and a choice test, where native speakers judge native texts (section 4).
- **Judgments of Non-Native Usage** As stated earlier, most computational work in the field of error detection tools for non-native speakers has relied on a single rater to annotate a gold standard corpus to check a system's output. We conduct an extensive double-annotation evaluation to measure inter-rater reliability and show that using one rater can be unreliable and may produce misleading results in a system test (section 5).
- **Sampling Approach** Multiple annotation can be very costly and time-consuming, which may explain why previous work employed only one rater. As an alternative to the standard exhaustive annotation, we propose a sampling approach in which estimates of the rates of hits, false positives, and misses are derived from random samples of the system's output, and then precision and recall of the system can be calculated. We show that estimates of system performance derived

[1] (Eeg-Olofsson and Knuttson, 2003) had a small evaluation of 40 prepositions and it is unclear whether they used multiple annotators or not.

[2] http://www.cambridge.org/elt

[3] http://langbank.engl.polyu.edu.hk/corpus/clec.html

from the sampling approach are comparable to those derived from an exhaustive annotation, but require only a fraction of the effort (section 6).

In short, through a battery of experiments we show how rating preposition usage, in either native or non-native texts, is a task that has surprisingly low inter-annotator reliability and thus greatly impacts system evaluation. We then describe a method for efficiently annotating non-native texts to make multiple annotation more feasible.

In section 2, we discuss in more depth the motivation for detecting usage errors in non-native writing, as well as the complexities of preposition usage. In section 3, we describe a system that automatically detects preposition errors involving incorrect selection and extraneous usage. In sections 4 and 5 respectively, we discuss experiments on the reliability of judging native and non-native preposition usage. In section 6, we present results of our system and results from comparing the sampling approach with the standard approach of exhaustive annotation.

2 Motivation

The long-term goal of our work is to develop a system which detects errors in grammar and usage so that appropriate feedback can be given to non-native English writers, a large and growing segment of the world's population. Estimates are that in China alone as many as 300 million people are currently studying English as a foreign language. Even in predominantly English-speaking countries, the proportion of non-native speakers can be very substantial. For example, the US National Center for Educational Statistics (2002) reported that nearly 10% of the students in the US public school population speak a language other than English and have limited English proficiency . At the university level in the US, there are estimated to be more than half a million foreign students whose native language is not English (Burghardt, 2002). Clearly, there is an increasing demand for tools for instruction in English as a Second Language (ESL).

Some of the most common types of ESL usage errors involve prepositions, determiners and collocations. In the work discussed here, we target preposition usage errors, specifically those of incorrect selection ("we arrived *to* the station") and extraneous use ("he went *to* outside")[4]. Preposition errors account for a substantial proportion of all ESL usage errors. For example, (Bitchener et al., 2005) found that preposition errors accounted for 29% of all the errors made by intermediate to advanced ESL students. In addition, such errors are relatively common. In our learner corpora, we found that 6% of all prepositions were incorrectly used. Some other estimates are even higher: for example, (Izumi et al., 2003) reported error rates that were as high as 10% in a Japanese learner corpus.

At least part of the difficulty in mastering prepositions seems to be due to the great variety of linguistic functions that they serve. When a preposition marks the argument of a predicate, such as a verb, an adjective, or a noun, preposition selection is constrained by the argument role that it marks, the noun which fills that role, and the particular predicate. Many English verbs also display alternations (Levin, 1993) in which an argument is sometimes marked by a preposition and sometimes not (e.g., "They loaded the wagon with hay" / "They loaded hay on the wagon"). When prepositions introduce adjuncts, such as those of time or manner, selection is constrained by the object of the preposition ("at length", "in time", "with haste"). Finally, the selection of a preposition for a given context also depends upon the intention of the writer ("we sat at the beach", "on the beach", "near the beach", "by the beach").

3 Automatically Detecting Preposition Usage Errors

In this section, we give a description of our system and compare its performance to other systems. Although the focus of this paper is on human judgments in the task of error detection, we describe our system to show that variability in human judgments can impact the evaluation of a system in this task. A full description of our system and its performance can be found in (Tetreault and Chodorow, 2008).

3.1 System

Our approach treats preposition error detection as a classification problem: that is, given a context of two words before and two words after the writer's preposition, what is the best preposition to use?

[4] There is a third error type, omission ("we are fond *null* beer"), that is a topic for our future research.

An error is marked when the system's suggestion differs from the writer's by a certain threshold amount.

We have used a maximum entropy (ME) classifier (Ratnaparkhi, 1998) to select the most probable preposition for a given context from a set of 34 common English prepositions. One advantage of using ME is that there are implementations of it which can handle very large models built from millions of training events and consisting of hundreds of thousands of feature-value pairs. To construct a model, we begin with a training corpus that is POS-tagged and heuristically chunked into noun phrases and verb phrases[5]. For each preposition that occurs in the training corpus, a preprocessing program extracts a total of 25 features. These consist of words and POS tags in positions adjacent to the preposition and in the heads of nearby phrases. In addition, we include combination features that merge the head features. We also include features representing only the tags to be able to cover cases in testing where the words in the context were not seen in training.

In many NLP tasks (parsing, POS-tagging, pronoun resolution), it is easy to acquire training data that is similar to the testing data. However, in the case of grammatical error detection, one does not have that luxury because reliable error-annotated ESL corpora that are large enough for training a statistical classifier simply do not exist. To circumvent this problem, we have trained our classifier on examples of prepositions used correctly, as in news text.

3.2 Evaluation

Before evaluating our system on non-native writing, we evaluated how well it does on the task of preposition selection in native text, an area where there has been relatively little work to date. In this task, the system predicts the writer's preposition based on its context. Its prediction is scored automatically by comparison to what the writer actually wrote. Most recently, (Gamon et al., 2008) addressed preposition selection by developing a system that combined a decision tree and a language model. Besides the difference in algorithms, there is also a difference in coverage between their system, which selects among 13 prepositions plus a category for *Other*, and the system presented here, which selects among 34 prepositions. In their system evaluation, they split a corpus of Reuters News text and Microsoft Encarta into two sets: 70% for training (3.2M examples), and the remaining 30% for testing (1.4M examples). For purposes of comparison, we used the same corpus and evaluation method. While (Gamon et al., 2008) do not present their overall accuracy figures on the Encarta evaluation, they do present the precision and recall scores for each preposition. In Table 3.2, we display their results in terms of F-measures and show the performance of our system for each preposition. Our model outperforms theirs for 9 out of the 10 prepositions that both systems handle. Overall accuracy for our system is 77.4% and increases to 79.0% when 7M more training examples are added. For comparison purposes, using a majority baseline (always selecting the preposition *of*) in this domain results in an accuracy of 27.2%.

Prep	(Gamon et al., 2008)	(Tetreault et al., 2008)
in	0.592	**0.845**
for	0.459	**0.698**
of	0.759	**0.906**
on	0.322	**0.751**
to	0.627	**0.775**
with	0.361	**0.675**
at	0.372	**0.685**
by	0.502	**0.747**
as	0.699	**0.711**
from	0.528	**0.591**
about	**0.800**	0.654

Table 1: Comparison of F-measures on Encarta/Reuters Corpus

(Felice and Pullman, 2007) used perceptron classifiers for preposition selection in BNC News Text at 85% accuracy. For each of the five most frequent prepositions, they used a separate binary classifier to decide whether that preposition should be used or not. The classifiers are not combined into a unified model. When we reconfigured our system and evaluation to be comparable to (Felice and Pullman, 2007), our model achieved an accuracy of 90% on the same five prepositions when tested on Wall Street Journal News, which is similar, though not identical, to BNC News.

While systems can perform at close to 80% accuracy in the task of preposition selection in native texts, this high performance does not transfer to the end-task of detecting preposition errors in essays by non-native writers. For example, (Izumi et al., 2003) reported precision and recall as low as 25% and 7% respectively when detecting different

[5] We have avoided parsing because our ultimate test corpus is non-native writing, text that is difficult to parse due to the presence of numerous errors in spelling and syntax.

grammar errors (one of which was prepositions) in English essays by non-native writers. (Gamon et al., 2008) reported precision up to 80% in their evaluation on the CLEC corpus, but no recall figure was reported. We have found that our system (the model which performs at 77.4%), also performs as high as 80% precision, but recall ranged from 12% to 26% depending on the non-native test corpus.

While our recall figures may seem low, especially when compared to other NLP tasks such as parsing and anaphora resolution, this is really a reflection of how difficult the task is. In addition, in error detection tasks, high precision (and thus low recall) is favored since one wants to minimize the number of false positives a student may see. This is a common practice in grammatical error detection applications, such as in (Han et al., 2006) and (Gamon et al., 2008).

4 Human Judgments of Native Usage

4.1 Cloze Test

With so many sources of variation in English preposition usage, we wondered if the task of selecting a preposition for a given context might prove challenging even for native speakers. To investigate this possibility, we randomly selected 200 sentences from Microsoft's Encarta Encyclopedia, and, in each sentence, we replaced a randomly selected preposition with a blank. We then asked two native English speakers to perform a cloze task by filling in the blank with the best preposition, given the context provided by the rest of the sentence. In addition, we had our system predict which preposition should fill each blank as well. Our results (Table 2) showed only about 76% agreement between the two raters (bottom row), and between 74% and 78% when each rater was compared individually with the original preposition used in Encarta. Surprisingly, the system performed just as well as the two native raters, when compared with Encarta (third row). Although these results seem very promising, it should be noted that in many cases where the system disagreed with Encarta, its prediction was not a good fit for the context. But in the cases where the raters disagreed with Encarta, their prepositions were also licensed by the context, and thus were acceptable alternatives to the preposition that was used in the text.

Our cloze study shows that even with well-formed text, native raters can disagree with each other by 25% in the task of preposition selection. We can expect even more disagreement when the task is preposition error detection in "noisy" learner texts.

	Agreement	Kappa
Encarta vs. Rater 1	0.78	0.73
Encarta vs. Rater 2	0.74	0.68
Encarta vs. System	0.75	0.68
Rater 1 vs. Rater 2	0.76	0.70

Table 2: Cloze Experiment on Encarta

4.2 Choice Test

The cloze test presented above was scored by automatically comparing the system's choice (or the rater's choice) with the preposition that was actually written. But there are many contexts that license multiple prepositions, and in these cases, requiring an exact match is too stringent a scoring criterion.

To investigate how the exact match metric might underestimate system performance, and to further test the reliability of human judgments in native text, we conducted a choice test in which two native English speakers were presented with 200 sentences from Encarta and were asked to select which of two prepositions better fit the context. One was the originally written preposition and the other was the system's suggestion, displayed in random order. The human raters were also given the option of marking both prepositions as equally good or equally bad. The results indicated that both Rater 1 and Rater 2 considered the system's preposition equal to or better than the writer's preposition in 28% of the cases. This suggests that 28% of the mismatched cases in the automatic evaluation are not system errors but rather are instances where the context licenses multiple prepositions. If these mismatches in the automatic evaluation are actually cases of correct system performance, then the Encarta/Reuters test which performs at 75% accuracy (third row of Table 2), is more realistically around 82% accuracy (28% of the 25% mismatch rate is 7%).

5 Annotator Reliability

In this section, we address the central problem of evaluating NLP error detection tools on learner data. As stated earlier, most previous work has relied on only one rater to either create an annotated corpus of learner errors, or to check the system's output. While some grammatical errors, such as number disagreement between subject and verb, no doubt show very high reliability, others, such as usage errors involving prepositions or determiners are likely to be much less reliable. In section 5.1, we describe our efforts in annotating a large corpus of student learner essays for preposition usage errors. Unlike previous work such as (Izumi et al., 2004) which required the rater to check for almost 40 different error types, we focus on annotating only preposition errors in hopes that having a single type of target will insure higher reliability by reducing the cognitive demands on the rater. Section 5.2 asks whether, under these conditions, one rater is acceptable for this task. In section 6, we describe an approach to efficiently evaluating a system that does not require the amount of effort needed in the standard approach to annotation.

5.1 Annotation Scheme

To create a gold-standard corpus of error annotations for system evaluation, and also to determine whether multiple raters are better than one, we trained two native English speakers to annotate preposition errors in ESL text. Both annotators had prior experience in NLP annotation and also in ESL error detection. The training was very extensive: both raters were trained on 2000 preposition contexts and the annotation manual was iteratively refined as necessary. To our knowledge, this is the first scheme that specifically targets annotating preposition errors[6].

The two raters were shown sentences randomly selected from student essays, with each preposition highlighted in the sentence. The raters were also shown the sentence which preceded the one containing the preposition that they rated. The annotator was first asked to indicate if there were any spelling errors within the context of the preposition (±2-word window and the commanding verb). Next the annotator noted determiner or plural errors in the context, and then checked if there were any other grammatical errors (for example, wrong verb form). The reason for having the annotators check spelling and grammar is that other modules in a grammatical error detection system would be responsible for these error types. For an example of a sentence with multiple spelling, grammatical and collocational errors, consider the following sentence: "In consion, for some reasons, museums, particuraly known travel place, get on many people." A spelling error follows the preposition *In*, and a collocational error surrounds *on*. If the contexts are not corrected, it is impossible to discern if the prepositions are correct. Of course, there is the chance that by removing these we will screen out cases where there are multiple interacting errors in the context that involve prepositions. When comparing human judgments to the performance of the preposition module, the latter should not be penalized for other kinds of errors in the context.

Finally, the annotator judged the writer's preposition with a rating of "0-extraneous preposition", "1-incorrect preposition", "2-correct preposition", or "e-equally good prepositions". If the writer used an incorrect preposition, the rater supplied the best preposition(s) given the context. Very often, when the writer's preposition was correct, several other prepositions could also have occurred in the same context. In these cases, the annotator was instructed to use the "e" category and list the other equally plausible alternatives. After judging the use of the preposition and, if applicable, supplying alternatives, the annotator indicated her confidence in her judgment on a 2-point scale of "1-low" and "2-high".

5.2 Two Raters vs. One?

Following training, each annotator judged approximately 18,000 occurrences of preposition use. Annotation of 500 occurrences took an average of 3 to 4 hours. In order to calculate agreement and kappa values, we periodically provided identical sets of 100 preposition occurrences for both annotators to judge (totaling 1800 in all). After removing instances where there were spelling or grammar errors, and after combining categories "2" and "e", both of which were judgments of correct usage, we computed the kappa values for the remaining doubly judged sets. These ranged from 0.411 to 0.786, with an overall combined value of 0.630[7]. The confusion matrix for the combined set (totaling 1336 contexts) is shown in Table 3. The rows represent Rater 1's (R1) judgments while the columns represent Rater 2's judgments. As one

[6] (Gamon et al., 2008) did not have a scheme for annotating preposition errors to create a gold standard corpus, but did use a scheme for the similar problem of verifying a system's output in preposition error detection.

[7] When including spelling and grammar annotations, kappa ranged from 0.474 to 0.773.

would expect given the prior reports of preposition error rates in non-native writing, the raters' agreement for this task was quite high overall (0.952) due primarily to the large agreement count where both annotators rated the usage "OK" (1213 total contexts). However there were 42 prepositions that both raters marked as a "Wrong Choice" and 17 as "Extraneous." It is important to note the disagreements in judging these errors: for example, Rater 1 judged 26 prepositions to be errors that Rater 2 judged to be OK, for a disagreement rate of .302 (26/86). Similarly, Rater 2 judged 37 prepositions to be errors that Rater 1 judged to be OK, for a disagreement rate of .381 (37/97).

R1↓; R2→	Extraneous	Wrong-Choice	OK
Extraneous	**17**	0	6
Wrong-Choice	1	**42**	20
OK	4	33	**1213**

Table 3: Confusion Matrix

The kappa of 0.630 and the off-diagonal cells in the confusion matrix both show the difficulty of this task and also show how two highly trained raters can produce very different judgments. This suggests that for certain error annotation tasks, such as preposition usage, it may not be appropriate to use only one rater and that using two or more raters to produce an adjudicated gold-standard set is the more acceptable path.

As a second test, we used a set of 2,000 preposition contexts from ESL essays (Chodorow et al., 2007) that were doubly annotated by native speakers with a scheme similar to that described above. We then compared an earlier version of our system to both raters' judgments, and found that there was a 10% difference in precision and a 5% difference in recall between the two system/rater comparisons. That means that if one is using only a single rater as a gold standard, there is the potential to over- or under-estimate precision by as much as 10%. Clearly this is problematic when evaluating a system's performance. The results are shown in Table 4.

	Precision	Recall
System vs. Rater 1	0.78	0.26
System vs. Rater 2	0.68	0.21

Table 4: Rater/System Comparison

6 Sampling Approach

If one uses multiple raters for error annotation, there is the possibility of creating an adjudicated set, or at least calculating the variability of system evaluation. However, annotation with multiple raters has its own disadvantages in that it is much more expensive and time-consuming. Even using one rater to produce a sizeable evaluation corpus of preposition errors is extremely costly. For example, if we assume that 500 prepositions can be annotated in 4 hours using our annotation scheme, and that the error rate for prepositions is 10%, then it would take at least 80 hours for a rater to find and mark 1000 errors. In this section, we propose a more efficient annotation approach to circumvent this problem.

6.1 Methodology

The sampling procedure outlined here is inspired by the one described in (Chodorow and Leacock, 2000). The central idea is to skew the annotation corpus so that it contains a greater proportion of errors. The result is that an annotator checks more potential errors since he or she is spending less time checking prepositions used correctly.

Here are the steps in the procedure. Figure 1 illustrates this procedure with a hypothetical corpus of 10,000 preposition examples.

1. Process a test corpus of sentences so that each preposition in the corpus is labeled "OK" or "Error" by the system.

2. Divide the processed corpus into two sub-corpora, one consisting of the system's "OK" prepositions and the other of the system's "Error" prepositions. For the hypothetical data in Figure 1, the "OK" sub-corpus contains 90% of the prepositions, and the "Error" sub-corpus contains the remaining 10%.

3. Randomly sample cases from each sub-corpus and combine the samples into an annotation set that is given to a "blind" human rater. We generally use a higher sampling rate for the "Error" sub-corpus because we want to "enrich" the annotation set with a larger proportion of errors than is found in the test corpus as a whole. In Figure 1, 75% of the "Error" sub-corpus is sampled while only 16% of the "OK" sub-corpus is sampled.

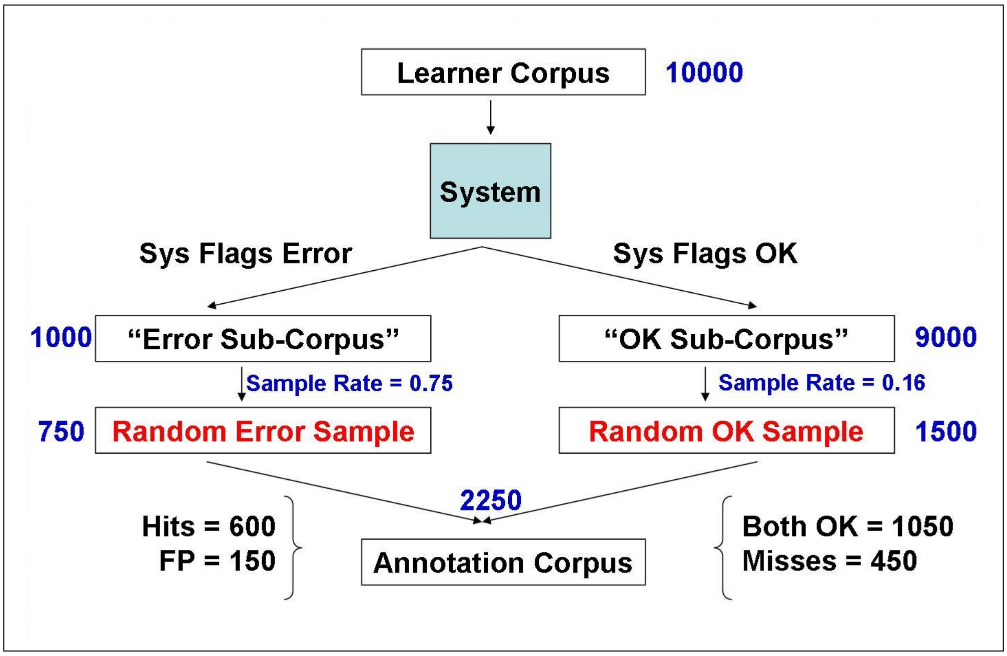

Figure 1: Sampling Approach (with hypothetical sample calculations)

4. For each case that the human rater judges to be an error, check to see which sub-corpus it came from. If it came from the "OK" sub-corpus, then the case is a Miss (an error that the system failed to detect). If it came from the "Error" sub-corpus, then the case is a Hit (an error that the system detected). If the rater judges a case to be a correct usage and it came from the "Error" sub-corpus, then it is a False Positive (FP).

5. Calculate the proportions of Hits and FPs in the sample from the "Error" sub-corpus. For the hypothetical data in Figure 1, these values are 600/750 = 0.80 for Hits, and 150/750 = 0.20 for FPs. Calculate the proportion of Misses in the sample from the "OK" sub-corpus. For the hypothetical data, this is 450/1500 = 0.30 for Misses.

6. The values computed in step 5 are conditional proportions based on the sub-corpora. To calculate the overall proportions in the test corpus, it is necessary to multiply each value by the relative size of its sub-corpus. This is shown in Table 5, where the proportion of Hits in the "Error" sub-corpus (0.80) is multiplied by the relative size of the "Error" sub-corpus (0.10) to produce an overall Hit rate (0.08). Overall rates for FPs and Misses are calculated in a similar manner.

7. Using the values from step 6, calculate Precision (Hits/(Hits + FP)) and Recall (Hits/(Hits + Misses)). These are shown in the last two rows of Table 5.

	Estimated Overall Rates Sample Proportion * Sub-Corpus Proportion
Hits	0.80 * 0.10 = **0.08**
FP	0.20 * 0.10 = **0.02**
Misses	0.30 * 0.90 = **0.27**
Precision	0.08/(0.08 + 0.02) = **0.80**
Recall	0.08/(0.08 + 0.27) = **0.23**

Table 5: Sampling Calculations (Hypothetical)

This method is similar in spirit to *active learning* ((Dagan and Engelson, 1995) and (Engelson and Dagan, 1996)), which has been used to iteratively build up an annotated corpus, but it differs from active learning applications in that there are no iterative loops between the system and the human annotator(s). In addition, while our methodology is used for *evaluating* a system, active learning is commonly used for *training* a system.

6.2 Application

Next, we tested whether our proposed sampling approach provides good estimates of a system's performance. For this task, we split a large corpus of ESL essays into two sets: first, a set of 8,269 preposition contexts (standard approach corpus) to be annotated using the scheme in section 5.1, and second, a set of 22,000 preposition contexts to be rated using the sampling approach (sampling corpus). We used two non-overlapping sets because the raters were the same for this test of the two approaches.

Using the standard approach, the sampling corpus of 22,000 prepositions would normally take several weeks for two raters to double annotate and then adjudicate. After this corpus was divided into "OK" and "Error" sub-corpora, the two sub-corpora were proportionally sampled, resulting in an annotation set of 750 preposition contexts (500 contexts from the "OK" sub-corpus and 250 contexts from the "Error" sub-corpus). This required roughly 6 hours for annotation, which is substantially more manageable than the standard approach. We had both raters work together to make judgments for each preposition context.

The precision and recall scores for both approaches are shown in Table 6 and are quite similar, thus suggesting that the sampling approach can be used as an alternative to exhaustive annotation.

	Precision	Recall
Standard Approach	0.80	0.12
Sampling Approach	0.79	0.14

Table 6: Sampling Results

6.3 Confidence Intervals

It is important with the sampling approach to use appropriate sample sizes when drawing from the sub-corpora, because the accuracy of the estimates of hits and misses will depend upon the proportion of errors in each sub-corpus as well as on the sample sizes. The "OK" sub-corpus is expected to have even fewer errors than the overall base rate, so it is especially important to have a relatively large sample from this sub-corpus. The comparison study described above used an "OK" sub-corpus sample that was twice as large as the Error sub-corpus sample.

One can compute the 95% confidence interval (CI) for the estimated rates of hits, misses and false positives by using the formula:

$$CI = p \pm 1.96 \times \sigma_p$$

where p is the proportion and σ_p is the standard error of the proportion given by:

$$\sigma_p = \sqrt{\frac{p(1-p)}{N}}$$

where N is the sample size.

For the example in Figure 1, the confidence interval for the proportion of Hits from the sample of the "Error" sub-corpus is:

$$CI_{hits} = 0.80 \pm 1.96 \times \sqrt{\frac{0.8 \times (1 - 0.80)}{750}}$$

which yields an interval of 0.077 and 0.083. Using these values, the confidence interval for precision is 0.77 to 0.83. The interval for recall can be computed in a similar manner. Of course, a larger sample size will yield narrower confidence intervals.

6.4 Summary

Table 7 summarizes the advantages and disadvantages of three methods for evaluating error detection systems. The standard (or exhaustive) approach refers to the method of annotating the errors in a large corpus. Its advantage is that the annotated corpus can be reused to evaluate the same system or compare multiple systems. However, it is costly and time-consuming which often precludes the use of multiple raters. The verification method (as used in (Gamon et al., 2008)), refers to the method of simply checking the acceptability of system output with respect to the writer's preposition. Like the sampling method, it has the advantages of efficiency and use of multiple raters (when compared to the standard method). But the disadvantage of verification is that it does not permit estimation of recall. Both verification and vampling methods require re-annotation for system retesting and comparison. In terms of system development, sampling (and to a lesser extent, verification) allows one to quickly assess system performance on a new corpus.

In short, the sampling approach is intended to alleviate the burden on annotators when faced with the task of having to rate several thousand errors of a particular type to produce a sizeable error corpus.

7 Conclusions

In this paper, we showed that the standard approach to evaluating NLP error detection systems (comparing the system's output with a gold-standard annotation) can greatly skew system results when the annotation is done by only one rater. However, one reason why a single rater is commonly used is that building a corpus of learner errors can be extremely costly and time-consuming. To address this efficiency issue, we presented a

Approach	Advantages	Disadvantages
Standard	Easy to retest system (no re-annotation required) Easy to compare systems Most reliably estimates precision and recall	Costly Time-Consuming Difficult to use multiple raters
Sampling	Efficient, especially for low-frequency errors Permits estimation of precision and recall More easily allows use of multiple raters	Less reliable estimate of recall Hard to re-test system (re-annotation required) Hard to compare systems
Verification	Efficient, especially for low-frequency errors More easily allows use of multiple raters	Does not permit estimation of recall Hard to re-test system (re-annotation required) Hard to compare systems

Table 7: Comparison of Evaluation Methods

sampling approach that produces results comparable to exhaustive annotation. This makes using multiple raters possible since less time is required to assess the system's performance. While the work presented here has focused on prepositions, the reasons for using multiple raters and a sampling approach apply equally to other error types, such as determiners and collocations.

It should be noted that the work here uses two raters. For future work, we plan on annotating preposition errors with more than two raters to derive a range of judgments. We also plan to look at the effects of feedback for errors involving prepositions and determiners, on the quality of ESL writing.

The preposition error detection system described here was recently integrated into *Criterion*$^{\text{SM}}$ Online Writing Evaluation Service developed by Educational Testing Service.

Acknowledgements We would first like to thank our two annotators Sarah Ohls and Waverly VanWinkle for their hours of hard work. We would also like to acknowledge the three anonymous reviewers and Derrick Higgins for their helpful comments and feedback.

References

Bitchener, J., S. Young, and D. Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*.

Burghardt, L. 2002. Foreign applications soar at universities. *New York Times*, April.

Chodorow, M. and C. Leacock. 2000. An unsupervised method for detecting grammatical errors. In *NAACL*.

Chodorow, M., J. Tetreault, and N-R. Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*.

Dagan, I. and S. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Proceedings of ICML*, pages 150–157.

Eeg-Olofsson, J. and O. Knuttson. 2003. Automatic grammar checking for second language learners - the use of prepositions. In *Nodalida*.

Engelson, S. and I. Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *Proceedings of ACL*, pages 319–326.

Felice, R. De and S. Pullman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*.

Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for esl error correction. In *IJCNLP*.

Han, N-R., M. Chodorow, and C. Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12:115–129.

Izumi, E., K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic error detection in the Japanese leaners' English spoken data. In *ACL*.

Izumi, E., K. Uchimoto, and H. Isahara. 2004. The overview of the sst speech corpus of Japanese learner English and evaluation through the experiment on automatic detection of learners' errors. In *LREC*.

Levin, B. 1993. *English verb classes and alternations: a preliminary investigation*. Univ. of Chicago Press.

Nagata, R., A. Kawai, K. Morihiro, and N. Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of the ACL/COLING*.

NCES. 2002. National center for educational statistics: Public school student counts, staff, and graduate counts by state: School year 2000-2001.

Ratnaparkhi, A. 1998. *Maximum Entropy Models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania.

Tetreault, J. and M. Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *COLING*.

Polysemy in verbs: systematic relations between senses and their effect on annotation

Anna Rumshisky
*Dept. of Computer Science
Brandeis University
Waltham, MA USA
arum@cs.brandeis.edu

Olga Batiukova†*
†Dept. of Spanish Philology
Madrid Autonomous University
Madrid, Spain
volha.batsiukova@uam.es

Abstract

Sense inventories for polysemous predicates are often comprised by a number of related senses. In this paper, we examine different types of relations within sense inventories and give a qualitative analysis of the effects they have on decisions made by the annotators and annotator error. We also discuss some common traps and pitfalls in design of sense inventories. We use the data set developed specifically for the task of annotating sense distinctions dependent predominantly on semantics of the arguments and only to a lesser extent on syntactic frame.

1 Introduction

Lexical ambiguity is pervasive in natural language, and its resolution has been used to improve performance of a number of natural language processing (NLP) applications, such as statistical machine translation (Chan et al., 2007; Carpuat and Wu, 2007), cross-language information retrieval and question answering (Resnik, 2006). Sense differentiation for the predicates depends on a number of factors, including syntactic frame, semantics of the arguments and adjuncts, contextual clues from the wider context, text domain identification, etc.

Preparing sense-tagged data for training and evaluation of word sense disambiguation (WSD) systems involves two stages: (1) creating a sense inventory and (2) applying it in annotation. Creating sense inventories for polysemous words is a task that is notoriously difficult to formalize. For polysemous verbs especially, constellations of related meanings make this task even more difficult. In lexicography, "lumping and splitting" senses during dictionary construction – i.e. deciding when to describe a set of usages as a separate sense – is a well-known problem (Hanks and Pustejovsky, 2005; Kilgarriff, 1997). It is often resolved on an ad-hoc basis, resulting in numerous cases of "overlapping senses", i.e. instances when the same occurrence may fall under more than one sense category simultaneously.

This problem has also been the subject of extensive study in lexical semantics, addressing questions such as when the context selects a distinct sense and when it merely modulates the meaning, what is the regular relationship between related senses, and what compositional processes are involved in sense selection (Pustejovsky, 1995; Cruse, 1995; Apresjan, 1973). A number of syntactic and semantic tests are traditionally applied for sense identification, such as examining synonym series, compatible syntactic environments, coordination tests such as *cross-understanding* or *zeugma* test (Cruse, 2000). None of these tests are conclusive and normally a combination of factors is used. At the recent Senseval competitions (Mihalcea et al., 2004; Snyder and Palmer, 2004; Preiss and Yarowsky, 2001), the choice of sense inventories frequently presented problems, spurring the efforts to create coarser-grained sense inventories (Hovy et al., 2006; Palmer et al., 2007; Navigli, 2006).

Part of the reason for such difficulties in establishing a set of senses available to a lexical item is that the meaning of a polysemous verb is often determined in composition and depends to the same extent on semantics of the particular arguments as it does on the base meaning of the verb itself. A number of systematic relations often holds between different senses of a polysemous verb. Depending on the kind of ambiguity involved in each case, some senses are easier to distinguish than others. Sense-tagged data (e.g. SemCor (Landes et al., 1998), PropBank (Palmer et al., 2005), OntoNotes (Hovy et al., 2006)) typically provides no way to differentiate between sense distinctions motivated by different factors. Treating different disambiguation factors separately would allow one to examine the contribution of each factor, as well as the success of a given algorithm in identifying the corresponding senses.

Within the scope of a sentence, syntactic frame and semantics of the arguments are most prominent in sense disambiguation. The latter is often more subtle and hence complex. Our goal in the present study was to target sense distinctions motivated strongly or exclusively by differences in argument semantics. We base the present discussion on the sense-tagged data set we developed for 20 polysemous verbs. We argue below that cases which can not be reliably disambiguated by humans introduce noise into the data and therefore should be kept out, a principle adhered to in the design of this data set.

The choice of argument semantics as the target disambiguation factor was motivated by several considerations. In automatic sense detection systems, argument semantics is often represented using external resources such as thesauri or shallow ontologies. Sense induction systems using distributional information often do not take into account the possible implications of induced word clusters for sense disambiguation. Our goal was to analyze differences in argument semantics that contribute to disambiguation.

In this paper, we discuss different kinds of systematic relations observed between senses of polysemous predicates and examine the effects they have on decisions made by the annotators. We also examine sense inventories for other factors that influence inter-annotator agreement rates and lead to annotation error. In Section 2, we discuss some of the factors that influence compilation of sense inventories and the methodology involved. In Section 3, we describe briefly the data set and the annotation task. In Sections 4 and 5, we discuss the relations observed between different senses within sense inventories in our data set, their effect on decisions made by the annotators, and the related annotation errors.

2 Defining A Sense Inventory

Several current resource-oriented projects undertake to formalize the procedure of identifying a word sense. FrameNet (Ruppenhofer et al., 2006) attempts to organize lexical information in terms of script-like semantic frames, with semantic and syntactic combinatorial possibilities specified for each frame-evoking lexical unit (word/sense pairing). Semantics of the arguments is represented by Fillmore's case roles (*frame elements*) which are derived on ad-hoc basis for each frame.

In OntoNotes project, annotators use small-scale corpus analysis to create sense inventories derived by grouping together WordNet senses. The procedure is restricted to maintain 90% inter-annotator agreement (Hovy et al., 2006).

Corpus Pattern Analysis (CPA) (Hanks and Pustejovsky, 2005; Pustejovsky et al., 2004) attempts to catalog prototypical norms of usage for individual words, specifying them in terms of context patterns. As a corpus analysis technique, CPA has its origins in the analysis of large corpora for lexicographic purposes, of the kind that was used for compiling the Cobuild dictionary (Sinclair and Hanks, 1987). Each pattern gives a combination of surface textual clues and argument specifications. A lexicographer creates a set of patterns by sorting a concordance for the target predicate according to the context features. In the present study, we use a modification of the CPA technique in the way explained in Section 3.

In CPA, syntactic and textual clues include argument structure and minor syntactic categories such as locatives and adjuncts; collocates from wider context; subphrasal cues such as genitives, partitives, bare plural/determiner, infinitivals, negatives, etc. Semantics of the arguments is represented either through a set of shallow semantic types corresponding to basic semantic features (e.g. Person, Location, PhysObj, Abstract, Event, etc.) or extensionally through *lexical sets*, which are effectively collections of lexical items.[1]

Several CPA patterns may correspond to a single sense. The patterns vary in syntactic structure or the encoding of semantic roles relative to the described event. For example, for the verb *treat*, DOCTOR treating PATIENT and DOCTOR treating DISEASE both correspond to the medical sense of *treat*. Knowing which semantic role is expressed by a particular argument is often useful for performing inference. For instance, treating a disease eliminates the disease, but not the patient. In the present annotation task, each pattern is viewed as **sense in construction** and labeled as a separate sense. In the rest of the paper, we will use the term "sense" to refer also to such microsenses.

For the cases where sense differentiation depends strongly on differences in semantics of the arguments, several factors further complicate creating a sense inventory. Prototypicality as a general principle of category organization seems to play an important role in defining both the boundaries of senses and the corresponding argument groupings. The same sense of the predicate is often activated by a number of semantically diverse arguments. Such argument sets are frequently organized around a core of typical members that are a "good fit" with respect to semantic requirements of the corresponding sense of the target. The relevant semantic feature is prominent for them, while other, more peripheral members of the argument set, merely allow the relevant interpretation (see Rumshisky (2008) for discussion). For example, the verb *absorb* has a sense involving *absorbing a substance*, and the typical members of the corresponding argument set would be actual substances, such as *oil, oxygen, water, air, salt*, etc. But *goodness, dirt, flavor, moisture* would also activate the same sense.

Each decision to split a sense and make another category is to a certain extent an arbitrary decision. For example, for the verb *absorb*, one can separate *absorbing a substance* (*oil, oxygen, water, air, salt*) from *absorbing energy* (*radiation, heat, sound, energy*). The latter sense may or may not be separated from *absorb-*

[1] See Rumshisky et al. (2006) and Pustejovsky et al. (2004) for more detail.

ing impact (*blow, shock, stress*). But it is a marked continuum, i.e. certain points in the continuum are more prominent, with necessity of a given concept reflected in the frequency of use.

When several senses are postulated based on argument distinctions, there are almost always *boundary cases* that can be seen to belong to both categories. Consider, for example, two senses defined for the verb *launch* and the corresponding direct objects in (1):

(1) a. *Physically propel an object into the air or water*
missile, rocket, torpedo, satellite, shuttle, craft
b. *Begin or initiate an endeavor*
campaign, initiative, investigation, expedition, drive, competition, crusade, attack, assault, inquiry

The senses seem to be very clearly separated, yet examples like *launch a ship* clearly fall on the boundary: while *ships* are physical objects propelled into water, *launching a ship* can be virtually synonymous with *launching an expedition*.

Similarly, for the verb *conclude*, two senses below which are linked to nominal complements are clearly separated:

(2) a. *finish*
meeting, debate, investigation, visit, tour, discussion; letter, chapter, novel
b. *reach an agreement*
treaty, agreement, deal, contract, truce, alliance, ceasefire, sale

However, *conclude negotiations* is clearly a boundary case where both interpretations are equally possible (negotiations may be concluded without reaching an agreement). In fact, the two annotators chose different senses for this example:[2]

(3) We were able to operate under a lease agreement until purchase negotiations were concluded.
annoA: *finish*
annoB: *reach an agreement*

In many cases, postulating a separate sense for a coherent set of nominal complements is not justified, as there are regular semantic processes that allow the complements to satisfy selectional requirements of the verb. For example, the verb *conclude*, in the *finish* sense accepts EVENT complements. Therefore, nouns such as *letter, chapter, novel* in (2) must be coerced into events corresponding to the activity that typically brings them about, that is, re-interpreted as events of writing (their Agentive quale, cf. Pustejovsky (1995)). Similarly, the verb *deny* in the first sense (*state or maintain that something is untrue*) accepts PROPOSITION complements:

(4) a. *state or maintain that something is untrue*
allegations, reports, rumour; significance, importance, difference; attack, assault, involvement
b. *refuse to grant something*
access, visa, approval, funding, license

Event nouns such as *attack* and *assault* are coerced into a propositional reading, as are relational nouns such as *significance* and *importance*.

Interestingly, as we have noted before (Rumshisky et al., 2006), each predicate imposes its own gradation with respect to prototypicality of elements of the argument set. As a result, even though basic semantic types such as PHYSOBJ, ANIMATE, EVENT, are used uniformly by many predicates, argument sets, while semantically similar, typically differ between predicates. For example, *fall* in the subject position and *cut* in the direct object position select for things that can be decreased:

(5) a. *cut (dobj)*: *reduce or lessen*
price, inflation, profits, cost, emission, spending, deficit, wages overhead, production, consumption, fees, staff
b. *fall (subj)*: *decrease*
price, inflation, profits, attendance, turnover, temperature, membership, import, demand, level

While there is a clear commonality between these argument sets, the overlap is only partial. To give another example, consider INFORMATION-selecting predicates *explain (subj)*, *grasp (dobj)* and *know (dobj)*. The nouns *book* and *note* occur in the subject position of *explain*; *answer* occurs both as the subject of *explain* and direct object of *know*; however, *grasp* accepts neither of these nouns as direct object. Thus, the actual selectional behavior of the predicates does not seem to be well described in terms of a fixed set of types, which is what is typically assumed by many ontologies used in automatic WSD.

3 Task Description

We were interested specifically in those cases where disambiguation needs to be made without relying on syntactic frame, and the main source of disambiguation is semantics of the arguments. Such cases are harder to identify formally in the development of sense inventories and harder for the annotators to determine. For example, phrasal verbs or idiomatic constructions that help identify a particular sense were intentionally excluded from our data set. Thus, for the verb *cut*, one of the senses involves cutting out a shape or a form (e.g. *cut a suit*), but the sentences with the corresponding phrasal form *cut out* were thrown out.

Even so, syntactic clues that contribute to disambiguation in some cases overrule the interpretation suggested by the argument. For example, for the verb *deny*, in *deny the attack*, the direct object strongly suggests a propositional interpretation for *deny* (that the attack didn't happen). However, the use of ditransitive construction (indicated in the example below by the past participle) overrules this interpretation, and we get the *refuse to grant* sense:

(6) Astorre, *denied* his *attack*, had stayed in camp, uneasily brooding.

[2] All examples are taken from the annotated data set. In some cases, sentence structure was slightly modified for brevity.

In fact, during the actual annotation, one of the annotators did not recognize the use of past participle, and erroneously assigned the *state or maintain something to be untrue* sense to this sentence.

3.1 Data set

The data set was developed using the British National Corpus (BNC), which is more balanced than the more commonly annotated Wall Street Journal data. We selected 20 polysemous verbs with sense distinctions that were judged to depend for disambiguation on semantics of the argument in several argument positions, including direct object (dobj), subject (subj), or indirect object within a prepositional phrase governed by *with* (iobj_with):

dobj: *absorb, acquire, admit, assume, claim, conclude, cut, deny, dictate, drive, edit, enjoy, fire, grasp, know, launch*
subj: *explain, fall, lead*
iobj_with: *meet*

We used the Sketch Engine (Kilgarriff et al., 2004) both to select the verbs and to aid the creation of the sense inventories. The Sketch Engine is a lexicographic tool that lists collocates that co-occur with a given target word in the specified grammatical relation. The collocates are sorted by their association score with the target.

A set of senses was created for each verb using a modification of the CPA technique (Pustejovsky et al., 2004). A set of complements was examined in the Sketch Engine. If a clear division was observed between semantically different groups of collocates in a certain argument position, the verb was selected. For semantically distinct groups of collocates, a separate sense was added to the sense inventory for the target. For example, for the verb *acquire*, a separate sense was added for each of the following sets of direct objects:

(7) a. *Take on certain characteristics*
shape, meaning, color, form, dimension, reality, significance, identity, appearance, characteristic, flavor
b. *Purchase or become the owner of property*
land, stock, business, property, wealth, subsidiary, estate, stake

The sense inventory for each verb was cross-checked against several resources, including WordNet, PropBank, Merriam-Webster and Oxford English dictionaries, and existing correspondences in FrameNet (Ruppenhofer et al., 2006; Hiroaki, 2003), OntoNotes (Hovy et al., 2006),[3] and CPA patterns (Hanks and Pustejovsky, 2005; Rumshisky and Pustejovsky, 2006; Pustejovsky et al., 2004).

We performed test annotation on 100 instances, with the sense inventory additionally modified upon examining the results of the annotation. This sense inventory was provided to two annotators, along with 200 sentences for each verb. Each sentence was pre-parsed with RASP (Briscoe and Carroll, 2002), and the head of the target argument phrase was identified. Misparses were manually corrected in post-processing.

3.2 Defining the task for the annotators

Data set creation for a WSD task is notoriously hard (cf. Palmer et al. (2007)), as the annotators are frequently forced to perform disambiguation on sentences where no disambiguation can really be performed. This is the case, for example, for overlapping senses, where more than one sense is activated simultaneously (Rumshisky, 2008; Pustejovsky and Boguraev, 1993). The goal was to create, for each target word, a set of instances where humans had no trouble disambiguating between different senses.

Two undergraduate linguistics majors served as annotators. The annotators were instructed to mark each sentence with the most fitting sense. The annotators were allowed to mark the sentence as “N/A” and were instructed to do so if (i) the sense inventory was missing the relevant sense, (ii) more than one sense seemed to fit, or (iii) the sense was impossible to determine from the context.

With respect to metaphoric senses, instructions were to throw out cases of creative use where the interpretation was difficult or not immediately clear. The cases where the target grammatical relation was actually absent from the sentence also had to be marked as “N/A” (e.g. for *fire*, sentences without direct object, e.g. *a stolen car was fired upon*). The annotators were also instructed to mark idiomatic expressions and phrasal verbs as “N/A”, e.g. for the verb *fall*: *fall from favor, fall through, fall in, fall back, fall silent, fall short, fall in love*.

Disagreements between the annotators were resolved in adjudication by the co-authors. The average inter-annotator agreement (ITA) for our data set was computed as a macro-average of the percentage of instances that were annotated with the same sense by both annotators to the total number of instances retained in the data set for each verb. The instances that were marked as “N/A” by one of the annotators (or thrown out during the adjudication) were not included in the computation. The ITA value for our data set was 95%. However, as we will see below, the ITA values do not always reflect the actual accuracy of annotation, due to some common problems with sense inventories.

3.3 Glossing a sense

A very common problem with glossing a sense involves the situation where a sense inventory includes two senses one of which is an extension of the other. The derived sense may be related to the primary sense through metaphor, and this often results in the former taking on a semantically less specific interpretation. The problem with creating glosses in this situation is that the words used may have sense distinctions

[3] Sense inventories released for the 65 verbs made available for SemEval-2007.

parallel to the ones in the target verb being described. This leaves the annotators free to choose either sense. This seems to be the case, for example, with OntoNotes sense inventory for *fire*, where *ignite or become ignited* is the gloss under which very divergent examples are grouped: *oil fired the furnace* (literal, primary sense) and *curiosity fired my imagination* (metaphoric extension). Clearly, annotators were having a problem with this sense due to the fact that the verb *ignite* has sense distinctions which are based on the same metaphor (*fire* = *inspire*) and therefore are very similar to those of the verb *fire*.

In case of semantic underspecification, annotators may be left free to choose the more generic sense, which contaminates the data set while not being reflected in the inter-annotator agreement values. For example, in our sense inventory for *acquire*, the gloss for *acquire a new customer* has to be very generic. We used the gloss "become associated with something, often newly brought into being". However, that led the annotators to overuse this gloss and select this sense in cases where a more specific gloss was more appropriate:[4]

(8) By this treaty, Russia *acquired* a Black Sea *coastline*.
annoA: *become associated with something, often newly brought into being*
annoB: *become associated with something, ...*
correct: *purchase or become the owner of property*

For a more detailed analysis of this phenomenon, see Section 5.

4 Relations Between Senses

In this section, we discuss linguistic processes underlying relations between senses within a single sense inventory. We believe that a detailed analysis of these processes should help to account for the annotator's ability to perform disambiguation. Some sense distinctions appear more striking to the annotators, depending on the type of relation involved.

In line with existing approaches to sense relations, we will look at both the linguistic structures involved in sense modification and the productive processes acting on linguistic structures. For the purposes of our present discussion, we interpret the literal (physical, direct) senses to be primary, with respect to more abstract or metaphorical senses.

4.1 Argument structure alternations

Some of the most striking differences between the senses are related to the argument structure alternations:

1. Different case roles (frame elements) may be expressed in the same argument position (in this case, direct object), corresponding to different perspectives on the same event. For example, direct object position of the verb *drive* may be filled by VEHICLE, DISTANCE, or PHYSOBJ giving rise to three distinct senses: (i) *operate a vehicle controlling its motion*, (ii) *travel in a vehicle a certain distance*, and (iii) *transport something or someone*. Similarly, for the verb *fire*, PROJECTILE or WEAPON in direct object position give rise to two related senses: (i) *shoot, discharge a weapon*, (ii) *shoot, propel a projectile*.

2. The distinction between propositional and non-propositional complements, as for the verbs *admit* and *deny* in (9) and (10):

(9) a. *admit defeat, inconsistency, offense*
(*acknowledge the truth or reality of*)
b. *admit patients, students*
(*grant entry or allow into a community*)

(10) a. *deny reports, importance, allegations*
(*state or maintain to be untrue*)
b. *deny visa, access*
(*refuse to grant*)

3. There is a mutual dependency between subcategorization features of the complements in different argument positions. For example, the [+animate] subject may combine with specific complements not available for [−animate], as for the two senses of *acquire*: (i) *learn* and (ii) *take on certain characteristics.* Compare NP$_{subj}$ [-animate] *acquire* NP$_{dobj}$ (*language, manners, knowledge, skill*) vs. NP$_{subj}$ [−animate] *acquire* NP$_{dobj}$ (*importance, significance*). Similarly, for *absorb*, compare NP$_{subj}$ [±animate] *absorb* NP$_{dobj}$ (*substance*) and NP$_{subj}$ [+animate] *absorb* NP$_{dobj}$ (*skill, information*). Note that, as one would expect, such dependencies are inevitable even despite the fact that our data set was developed specifically to target sense distinctions dependent on a single argument position.

4.2 Event structure modification

Event structure modifications (i.e. operations affecting aspectual properties of the predicate) are another source of sense differentiation. Two cases appear most prominent:

1. The event structure is modified along with the characteristics of the arguments. For example, for *enjoy*, compare *enjoy skiing, vacation* (DYNAMIC EVENT) with *enjoying a status* (STATE). Similarly, for *lead*, compare *a person leads smb somewhere* (PROCESS) vs. *a road* (PATH) *leads somewhere* (STATE); for *explain*, compare *something or somebody explains smth* (= *clarifies, describes, makes comprehensible*, PROCESS) vs. *something* [−inanimate, +abstract] *explains something* (= *is a reason for something*, STATE); for *fall*, compare PHYSOBJ *falls* (TRANSITION or ACCOMPLISHMENT) vs. *a case falls into a certain category* (STATE).

2. The aspectual nature of the predicate is the only semantically relevant feature that remains unchanged after consecutive sense modifications. For example, the ingressive meaning of 'beginning something' is preserved in shifting from the physical sense of the verb *launch* in *launch a missile* to *launch a campaign* and *launch a product*.

[4] We will refer to annotators A and B as *annoA* and *annoB*.

4.3 Lexical semantic features

Sense distinctions often involve deeper semantic characteristics of the verbs which could be accounted for by means of lexical semantic features such as qualia structure roles in Generative Lexicon (Pustejovsky, 1995):[5]

1. Consider how the meaning component 'manner of motion' (typically associated with the agentive role) gets transformed in the different senses of *drive*. It is obviously present in the physical uses of *drive* (such as *operate a vehicle, transport something or somebody*, etc.), but is completely lost in *motivate the progress of* (as in *drive the economy, drive the market forward*, etc.). The value of the agentive role of *drive* becomes underspecified or semantically weak, so that the overall meaning of *drive* is transformed to *cause something to move.*

2. Information about semantic type contained in qualia structure allows apparently diverse elements to activate the same sense of the verb. For instance, the verb *absorb* in the sense *learn or incorporate skill or information* occurs with direct objects such as *values, atmosphere, information, idea, words, lesson, attitudes, culture*. The requisite semantic component is realized differently for each of these words. Some of them are complex types[6] with INFORMATION as one of the constituent types: *words* (ACOUSTIC/VISUAL ENTITY • INFO), *lesson* (EVENT • INFO). Others, such as *idea*, are polysemous, with one of the senses being INFORMATION. Cases like *culture* and *values* are more difficult, but since they refer to knowledge, the INFORMATION component is clearly present. Consequently, the annotators are able to identify the corresponding sense of *absorb* with a high degree of agreement.

4.4 Metaphor and metonymy

The processes causing the mentioned meaning transformations in our corpus often involve metaphor and metonymy. Below are some of the conventionalized extensions with metaphorical flavor:

(11) a. *grasp object* vs. *grasp meaning*
b. *launch object* vs. *launch an event (campaign, assault)* or *launch a product (newspaper, collection)*
c. *meet with a person* vs. *meet with success, resistance*
d. *lead somebody somewhere* vs. *lead to a consequence*

Note that these metaphorical extensions involve abstract or continuous objects (*meaning, assault, success, consequence*), which in turn cause event structure modifications (*lead* as a process vs. *lead* as a state). Thus, the processes and structures we are dealing with are clearly interrelated.

The metonymical process can be exemplified by *edit* as *make changes to the text* and as *supervise publication*, which are in a clear contiguity relationship.

One of the effects of the metaphorization and progressive emptying of the primary (physical, concrete) senses is the distinction between generic and specific senses. For example, compare *acquire land, business* (specific sense) to *acquire an infection, a boyfriend, a following,* which refers to some extremely light generic association. Similar process is observed for the semantically weak sense of *fall, be associated with or get assigned to a person or location or for event to fall onto a time*:

(12) Birthdays, lunches, celebrations *fall* on a certain date or time
Stress or emphasis *fall* on a given topic or a syllable
Responsibility, luck, suspicion *fall* on or to a person

The specificity often involves specialization within a certain domain:

(13) a. *conclude* as *finish* vs. *conclude* as *reach an agreement* (Law, Politics)
b. *fire* as *shoot a weapon or a projectile* vs. *fire* as *kick or pass an object of play in sports* (Sport)

Thus, when concluding a *pact* or an *agreement*, a certain EVENT is also being finished (negotiation of that agreement), necessarily with a positive outcome.

In the following section, we will try to show how different kinds of relations between senses influence disambiguation carried out by the annotators. In particular, we look at different sources of disagreement and annotator error as determined in adjudication.

5 Analysis of Annotation Decisions

As we have seen above, in many cases disambiguation is impossible due to the nature of compositionality. Also, as there are no clear answers to a number of questions concerning sense identification, the annotators deal with sense inventories that are imperfect. Results of the disambiguation task carried out by the annotators reflect all these defects.

In cases when a specific meaning from the data set is not included into the sense inventory (e.g. due to its low frequency or extreme fine-grainedness) the annotators may use a more general meaning or pick the closest meaning available. For example, within the sense inventory for *fire*, there was no separate gloss for *fire an engine.* Annotator A in our experiment chose the closest specific meaning available, and Annotator B marked it with a more generic sense:

(14) Engineers successfully *fired* thrusters to boost the research satellite to an altitude of 507 km.
annoA: *shoot, propel a projectile*
annoB: *apply fire to*

As mentioned in Section 3.3, even when the appropriate specific sense is available, annotators frequently chose the more generic sense in its place, as in (15), (16) and (17), and also in (8).

[5] We will use the terminology from Generative Lexicon (Pustejovsky, 1995; Pustejovsky, 2007) to discuss lexical semantic properties, such as *qualia roles, complex* and *functional types*, and so on.

[6] *Complex type* is a term used for concepts that inherently refer to more than one semantic type.

(15) Several *referrals fell* into this *category*.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *be categorized as or fall into a range*

(16) The terrible *silence* had *fallen*.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *for a state (such as darkness or silence) to come, to commence*

(17) He *acquired* a *taste* for performing in public.
annoA: *become associated with something, often newly brought into being*
annoB: *become associated with something, ...*
correct: *learn*

Note that in (8) this decision was probably motivated by the annotators' uncertainty about the semantic ascription of the relevant argument (*coastline* is not a prototypical owned property). The generic sense seems to be the safest option to take for the annotators, as compared to taking a chance with a specific meaning. Due to its low degree of semantic specification, the generic sense is potentially able to embrace almost every possible use. This is not a desirable outcome because the generic senses are introduced in the inventory to account only for semantically underspecified cases. For instance, *become associated with something, often newly brought into being* is appropriate for *acquire a grandchild*, but not for *acquire a taste* or *acquire a proficiency*.

Remarkable variation is also observed with respect to **non-literal uses** as discussed in Section 4.4. For example, in (18) and (19) abstract NPs *panic* and *imbalance of forces* are equated with *energy or impact* by one annotator and with *substance* by the other.

(18) Her *panic* was *absorbed* by his warmth.
annoA: *absorb energy or impact*
annoB: *absorb substance*

(19) Alternatively, *imbalance* of forces can be *absorbed* into the body.
annoA: *absorb energy or impact*
annoB: *absorb substance*

In some cases, the literal and the metaphoric senses are activated simultaneously resulting in ambiguity (cf. Cruse (2000)):

(20) For over 300 years this waterfall has provided the energy to *drive* the *wheels* of industry.
annoA: *motivate the progress of*
annoB: *provide power for or physically move a mechanism*

(21) But fashion changed and the short *skirt fell* – literally – from favour and started skimming the ankles.
annoA: *lose power or suffer a defeat*
annoB: N/A

(22) She was delighted when the *story* of Hank *fell* into her lap.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *physically drop; move or extend downward*

Impact of **subcategorization features** on disambiguation (cf. Section 4.1 para 3) is illustrated in (23).

(23) The reggae tourist can easily *absorb* the current reggae *vibe*.
annoA: *absorb energy or impact*
annoB: *learn or incorporate skill or information*

Both interpretations chosen here (*absorb energy or impact* and *learn or incorporate skill or information*) were possible due to the animacy of the subject, which activates two different subcategorization frames and subsequently two different senses.

Typically, cases where **semantic type** of the relevant arguments (cf. Section 4.3 para 2) is not clear result in annotator disagreement:

(24) The AAA *launched* education *programs*.
annoA: *begin or initiate an endeavor* (EVENT)
annoB: *begin to produce or distribute; start a company* (PRODUCT)

(25) France plans to *launch* a remote-sensing *vehicle* called Spot.
annoA: *physically propel into the air, water or space* (PHYSOBJ)
annoB: *begin to produce or distribute; start a company* (PRODUCT)

The two cases above are interesting in that both *program* and *vehicle* are ambiguous and can be analyzed semantically as members of different semantic classes. This is what the annotators in fact do, and as a result, ascribe them to different senses. *Program* can be categorized as EVENT ('series of steps') or as INTELLECTUAL ACTIVITY PRODUCT ('document or system of projects'). It is a complex type, i.e. it is an inherently polysemous word that represents at least two different semantic types. *Vehicle*, in turn, is a functional type: on the one hand, it represents an entity with certain formal properties (PHYSOBJ interpretation), on the other hand, it is an artifact, with a prominent practical purpose (PRODUCT interpretation).

In fact, most problems the annotators had with the task are due to the inherent semantic complexity of words such as *vehicle* and *program* in (24) and (25) and to the existence of boundary cases, where the relevant noun does not properly belong to one or another semantic category. This is the case with *panic*, *imbalance* or *reggae vibe* in (18), (19), and (23), and also with *taste* and *coastline* in (17) and (7).

In some of these cases, other contextual clues may come into play and tip the balance in favor of one or another sense. Note that disambiguation was influenced by a **wider context** even despite the intentionally restrictive task design (targeting a particular syntactic relation for each verb). For instance, in (26), **domain-specific clues** referring to war or military conflict (such as *rebel control*) could have motivated Annotator B's decision to ascribe it to the sense *lose power or suffer a defeat* (even though a road is not typically an entity that can lose power), while the other annotator chose a more generic meaning:

(26) The *road fell* into rebel control.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *lose power or suffer a defeat*

Other pragmatic and discourse-oriented clues played a role, in particular, positive and negative connotation of the senses and the relevant arguments, as well as the temporal organization of discourse. For example, in (27) and (28), positive or neutral interpretation of *wave of immigrants* and *change* could have led to the choice of *take in or assimilate* and *learn or incorporate skill or information* senses, while the negatively-colored interpretation might explain the choice of the *bear the cost of* sense.

(27) ..help *absorb* the latest *wave of immigrants*.
annoA: *bear the cost of; take on an expense*
annoB: *take in or assimilate, making part of a whole or a group*

(28) For senior management an important lesson was the trade unions' capacity to *absorb change* and to become its agents.
annoA: *learn or incorporate skill or information*
annoB: *bear the cost of; take on an expense*

Temporal organization of a broader discourse is another important factor. For example, for the verb *claim*, the senses *claim the truth of* and *claim property you are entitled to* have different presuppositions with respect to preexistence of the thing claimed. In (28), due to the absence of a broader context, the annotators chose two different temporal reference interpretations. For Annotator B, *success* was something that has happened already, while for A this was not clear (*success* might have been achieved or not):

(29) One area where the government can *claim* some *success* involves debt repayment.
annoA: *come in possession of or claim property you are entitled to*
annoB: *claim the truth of*

6 Conclusion

We have given a brief overview of different types of sense relations commonly found in polysemous predicates and analyzed their effect on different aspects of the annotation task, including sense inventory design and execution of the WSD annotation.

The present analysis suggests that theoretical tools must be refined and further developed in order to give an adequate account to the sense modifications found in real corpus data. To this end, broader contextual clues and discourse-oriented clues need to be included in the analysis.

Semantically annotated corpora are routinely developed for the training and testing of automatic sense detection and induction algorithms. But they do not typically provide a way to distinguish between different kinds of ambiguities. Consequently, it is difficult to perform adequate error analysis for different sense detection systems. Appropriate semantic annotation that would allow one to determine which sense distinctions can be detected better by automatic systems does not need to be highly specific and unnecessarily complex, but requires development of robust generalizations about sense relations.

One obvious conclusion is that data sets need to be explicitly restricted to the instances where humans have no trouble disambiguating between different senses. Thus, prototypical cases can be accounted for reliably, ensuring the clarity of annotated sense distinctions. At face value, imposing such restrictions may appear to negatively influence usability of the resulting data set in particular applications requiring WSD, such as machine translation or information retrieval. However, this decision impacts most strongly those boundary cases which are not reliably disambiguated by human annotators, and which rather introduce noise into the data set.

Acknowledgments

This work was supported in part by NSF CRI grant to Brandeis University. The work of O. Batiukova is supported by postdoctoral grant of the Ministry of Education of Spain and Madrid Autonomous University.

References

Apresjan, Ju. 1973. Regular polysemy. *Linguistics*, 142(5):5–32.

Briscoe, T. and J. Carroll. 2002. Robust accurate statistical annotation of general text. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, May 2002*, pages 1499–1504.

Carpuat, M. and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proc. of EMNLP-CoNLL*, pages 61–72.

Chan, Y. S., H. T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL*, pages 33–40, Prague, Czech Republic, June.

Cruse, D. A. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Dizier, Patrick St. and Evelyne Viegas, editors, *Computational Lexical Semantics*, pages 33–49. Cambridge University Press, Cambridge, England.

Cruse, D. A. 2000. *Meaning in Language, an Introduction to Semantics and Pragmatics*. Oxford University Press, Oxford, United Kingdom.

Hanks, P. and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*.

Hiroaki, S. 2003. FrameSQL: A software tool for FrameNet. In *Proceedigns of ASIALEX '03*, pages 251–258, Tokyo, Japan. Asian Association of Lexicography.

Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Kilgarriff, A., P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

Kilgarriff, A. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

Landes, S., C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In Fellbaum, C., editor, *Wordnet: an electronic lexical database*. MIT Press, Cambridge (Mass.).

Mihalcea, R., T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In Mihalcea, Rada and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain, July. Association for Computational Linguistics.

Navigli, R. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia, July. Association for Computational Linguistics.

Palmer, M., D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, M., H. Dang, and C. Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

Preiss, J and D. Yarowsky, editors. 2001. *Proceedings of the Second Int. Workshop on Evaluating WSD Systems (Senseval 2)*. ACL2002/EACL2001.

Pustejovsky, J. and B. Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artif. Intell.*, 63(1-2):193–223.

Pustejovsky, J., P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

Pustejovsky, J. 1995. *Generative Lexicon*. Cambridge (Mass.): MIT Press.

Pustejovsky, J. 2007. Type Theory and Lexical Decomposition. In Bouillon, P. and C. Lee, editors, *Trends in Generative Lexicon Theory*. Kluwer Publishers (in press).

Resnik, P. 2006. Word sense disambiguation in NLP applications. In Agirre, E. and P. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Rumshisky, A. and J. Pustejovsky. 2006. Inducing sense-discriminating context patterns from sense-tagged corpora. In *LREC 2006, Genoa, Italy*.

Rumshisky, A., P. Hanks, C. Havasi, and J. Pustejovsky. 2006. Constructing a corpus-based ontology using model bias. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA.

Rumshisky, A. 2008. Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica*. forthcoming.

Ruppenhofer, J., M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

Sinclair, J. and P. Hanks. 1987. *The Collins Cobuild English Language Dictionary*. HarperCollins, 4th edition (2003) edition. Published as Collins Cobuild Advanced Learner's English Dictionary.

Snyder, B. and M. Palmer. 2004. The english all-words task. In Mihalcea, Rada and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Eliciting Subjectivity and Polarity Judgements on Word Senses

Fangzhong Su
School of Computing
University of Leeds
fzsu@comp.leeds.ac.uk

Katja Markert
School of Computing
University of Leeds
markert@comp.leeds.ac.uk

Abstract

There has been extensive work on eliciting human judgements on the sentiment of words and the resulting annotated word lists have frequently been used for opinion mining applications in Natural Language Processing (NLP). However, this word-based approach does not take different senses of a word into account, which might differ in whether and what kind of sentiment they evoke. In this paper, we therefore introduce a human annotation scheme for judging both the subjectivity and polarity of *word senses*. We show that the scheme is overall reliable, making this a well-defined task for automatic processing. We also discuss three issues that surfaced during annotation: the role of annotation bias, hierarchical annotation (or underspecification) and bias in the sense inventory used.

1 Introduction

Work in psychology, linguistics and computational linguistics has explored the affective connotations of words via eliciting human judgements (see Section 2 for an in-depth review). Two important parameters in determining affective meaning that have emerged are subjectivity and polarity. *Subjectivity identification* focuses on determining whether a language unit (such as a word, sentence or document) is *subjective*, i.e. whether it expresses a *private state, opinion or attitude*, or is factual. *Polarity identification* focuses on whether a language unit has a positive or negative connotation.

Word lists that result from such studies would, for example tag *good* or *positive* as a positive word, *bad* as negative and *table* as neither. Such word lists have frequently been used in natural language processing applications, such as the automatic identification of a review as favourable or unfavourable (Das and Chen, 2001). However, the word-based annotation conducted so far is at least partially unreliable. Thus Andreevskaia and Bergler (2006) find only a 78.7% agreement on subjectivity/polarity tags between two widely used word lists. One problem they identify is that word-based annotation does not take different senses of a word into account. Thus, many words are *subjectivity-ambiguous* or *polarity-ambiguous*, i.e. have both subjective and objective or both positive and negative senses, such as the words *positive* and *catch* with corresponding example senses given below.[1]

(1) positive, electropositive—having a positive electric charge;“protons are positive” (*objective*)

(2) plus, positive—involving advantage or good; “a plus (or positive) factor” (*subjective*)

(3) catch—a hidden drawback; “it sounds good but what’s the catch?” (*negative*)

(4) catch, match—a person regarded as a good matrimonial prospect (*positive*)

Inspired by Andreeivskaia and Bergler (2006) and Wiebe and Mihalcea (2006), we therefore explore the subjectivity and polarity annotation of *word senses* instead of *words*. We hypothesize that annotation at the sense level might eliminate one possible source of disagreement for subjectivity/polarity annotation and will therefore hopefully lead to higher agreement than at the word level.

[1] All examples in this paper are from WordNet 2.0.

An additional advantage for practical purposes is that subjectivity labels for senses add an additional layer of annotation to electronic lexica and can therefore increase their usability. As an example, Wiebe and Mihalcea (2006) prove that subjectivity information for WordNet senses can improve word sense disambiguation tasks for subjectivity-ambiguous words (such as *positive*). In addition, Andreevskaia and Bergler (2006) show that the performance of automatic annotation of subjectivity at the *word* level can be hurt by the presence of subjectivity-ambiguous words in the training sets they use. A potential disadvantage for annotation at the sense level is that it is dependent on a lexical resource for sense distinctions and that an annotation scheme might have to take idiosyncracies of specific resources into account or, ideally, abstract away from them.

In this paper, we investigate the reliability of manual subjectivity labeling of word senses. Specifically, we mark up subjectivity/attitude (subjective, objective, and both) of word senses as well as polarity/connotation (positive, negative and no polarity). To the best of our knowledge, this is the first annotation scheme for judging both subjectivity and polarity of word senses. We test its reliability on the WordNet sense inventory. Overall, the experimental results show high agreement, confirming our hypothesis that agreement at sense level might be higher than at the word level. The annotated sense inventory will be made publically available to other researchers at `http://www.comp.leeds.ac.uk/markert/data`.

The remainder of this paper is organized as follows. Section 2 discusses previous related work. Section 3 describes our human annotation scheme for word sense subjectivity and polarity in detail. Section 4 presents the experimental results and evaluation. We also discuss the problems of bias in the annotation scheme, the impact of hierarchical organization or underspecification on agreement as well as problems with bias in WordNet sense descriptions. Section 5 compares our annotation to the annotation of a different scheme, followed by conclusions and future work in Section 6.

2 Related Work

Osgood et al. (1957) proposed semantic differential to measure the connotative meaning of concepts. They conducted a factor analysis of large collections of semantic differential scales and pointed out three referring attitudes that people use to evaluate words and phrases—evaluation (good-bad), potency (strong-weak), and activity (active-passive). Also, they showed that these three dimensions of affective meaning are cross-cultural universals from a study on dozens of cultures (Osgood et al., 1975). This work has spawned a considerable amount of linguistic and psychological work in affect analysis on the *word* level. In psychology both the **Affective Norms for English Words (ANEW)** project as well as the **Magellan** project focus on collecting human judgements on affective meanings of words, roughly following Osgood's scheme. In the ANEW project they collected numerical ratings of pleasure (equivalent to our term polarity), arousal, and dominance for 1000 English terms (Bradley and Lang, 2006) and in Magellan they collected cross-cultural affective meanings (including polarity) in a wide variety of countries such as the USA, China, Japan, and Germany (Heise, 2001). Both projects concentrate on collecting a large number of ratings on a large variety of words: there is no principled evaluation of agreement.

The more linguistically oriented projects of the **General Inquirer (GI)** lexicon[2] and the **Appraisal** framework [3] also provide word lists annotated for affective meanings but judgements seem to be currently provided by one researcher only. Especially the General Enquirer which contains 11788 words marked for polarity (1915 positive, 2291 negative and 7582 no-polarity words) seems to use a relatively ad hoc definition of polarity. Thus, for example *amelioration* is marked as no-polarity whereas *improvement* is marked as positive.

The projects mentioned above center on subjectivity analysis on words and therefore are not good at dealing with subjectivity or polarity-ambiguous words as explained in the Introduction. Work that like us concentrates on *word senses* includes approaches where the subjectivity labels are automatically assigned such as **WordNet-Affect** (Strapparava and Valitutti, 2004), which is a subset of WordNet senses with semi-automatically assigned affective labels (such as emotion, mood or behaviour). In a first step, they manually collect an affective word list and a list of synsets which contain at least one word in this word list. Fine-

[2] Available at http://www.wjh.harvard.edu/ inquirer/

[3] Available at http://www.grammatics.com/appraisal/

grained affect labels are assigned to these synsets by the resource developers. Then they automatically expand the lists by employing WordNet relations which they consider to reliably preserve the involved labels (such as similar-to, antonym, derived-from, pertains-to, and attribute). Our work differs from theirs in three respects. First, they focus on their semi-automatic procedure, whereas we are interested in *human judgements*. Second, they use a finer-grained set of affect labels. Third, they do not provide agreement results for their annotation. Similarly, **SentiWordNet**[4] is a resource with *automatically determined polarity* of word senses in WordNet (Esuli and Sebastiani, 2006), produced via bootstrapping from a small manually determined seed set. Each synset has three scores assigned, representing the positive, negative and neutral score respectively. No human annotation study is conducted.

There are only two human annotation studies on subjectivity of word senses as far as we are aware. Firstly, the **Micro-WNOp** corpus is a list of about 1000 WordNet synsets annotated by Cerini et al. (2007) for polarity. The raters manually assigned a triplet of numerical scores to each sense which represent the strength of positivity, negativity, and neutrality respectively. Their work differs from us in two main aspects. First, they focus on polarity instead of subjectivity annotation (see Section 3 for a discussion of the two concepts). Second, they do not use absolute categories but give a rating between 0 and 1 to each synset—thus a synset could have a non-zero rating on both negativity and positivity. They also do not report on agreement results. Secondly, **Wiebe and Mihalcea (2006)** mark up WordNet senses as subjective, objective or both with good agreement. However, we expand their annotation scheme with polarity annotation. In addition, we hope to annotate a larger set of word senses.

3 Human Judgements on Word Sense Subjectivity and Polarity

We follow Wiebe and Mihalcea (2006) in that we see subjective expressions as private states "that are not open to objective observation or verification" and in that annotators distinguish between subjective (*S*), objective (*O*) and both subjective/objective (*B*) senses.

Polarity refers to positive or negative connotations associated with a word or sense. In contrast to other researchers (Hatzivassiloglou and McKeown, 1997; Takamura et al., 2005), we do not see polarity as a category that is dependent on prior subjectivity assignment and therefore applicable to subjective senses only. Whereas there is a dependency in that most subjective senses have a relatively clear polarity, polarity can be attached to objective words/senses as well. For example, *tuberculosis* is not subjective — it does not describe a private state, is objectively verifiable and would not cause a sentence containing it to carry an opinion, but it does carry negative associations for the vast majority of people. We allow for the polarity categories positive (*P*), negative (*N*), varying (*V*) or no-polarity (*NoPol*).

Overall we combine these annotations into 7 categories—*S:N*, *S:P*, *S:V*, *B*, *O:N*, *O:P*, and *O:NoPol*, which are explained in detail in the subsequent sections. Figure 1 gives an overview of the hierarchies over all categories.

As can be seen in Figure 1, our annotation scheme allows for hierarchical annotation, i.e. it is possible to only annotate for subjectivity or polarity. This can be necessary to achieve higher agreement by merging categories or to concentrate in specific applications on only one aspect.

3.1 Subjectivity

3.1.1 Subjective Senses

Subjective senses include several categories, which can be expressed by nouns, verbs, adjectives or adverbs. Firstly, we include emotions. Secondly, we include judgements, assessments and evaluations of behaviour as well as aesthetic assessments of individuals, natural objects and artefacts. Thirdly, mental states such as doubts, beliefs and speculations are also covered by our definition. This grouping follows relatively closely the definition of attitudinal positioning in the Appraisal scheme (which has, however, only been used on words, not on word senses before).

These types of subjectivity can be expressed via direct references to an emotion or mental state (see Example 5 or 8 below) as well as by expressive subjective elements (Wiebe and Mihalcea, 2006). Expressive subjective elements contain judgemental references to objects or events. Thus, *pontificate* in Example 6 below is a reference to a speech event that always judges it negatively; *beautiful* as

[4] Available at http://sentiwordnet.isti.cnr.it/

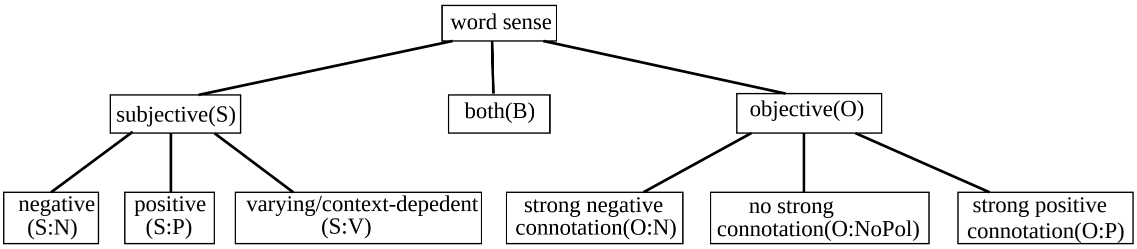
word sense
subjective(S)
both(B)
objective(O)
negative (S:N)
positive (S:P)
varying/context-depedent (S:V)
strong negative connotation(O:N)
no strong connotation(O:NoPol)
strong positive connotation(O:P)

Figure 1: Overview of the hierarchies over all categories

in Example 7 below is a positive judgement.

(5) angry—feeling or showing anger; "angry at the weather; "angry customers; an angry silence" (*emotion*)

(6) pontificate—talk in a dogmatic and pompous manner; "The new professor always pontificates" (*assessment of behaviour*)

(7) beautiful—aesthetically pleasing (*aesthetic assessment*)

(8) doubt, uncertainty, incertitude, dubiety, doubtfulness, dubiousness—the state of being unsure of something (*mental state*)

3.1.2 Objective Senses

Objective senses refer to persons, objects, actions, events or states without an inherent emotion or judgement or an expression of a mental state. Examples are references to individuals via named entities (see Example 9) or non-judgemental references to artefacts, persons, animals, plants, states or events (see Example 10 and 11). If a sentence contains an opinion, it is not normally due to the presence of this word sense and the sense often expresses objectively verifiable states or events. Thus, Example 12 is objective as we can verify whether there is a war going on. In addition, a sentence containing this sense of *war* does not necessarily express an opinion.

(9) Einstein, Albert Einstein – physicist born in Germany who formulated the special theory of relativity and the general theory of relativity; Einstein also proposed that light consists of discrete quantized bundles of energy (later called photons) (1879-1955) (*named entity*)

(10) lawyer, attorney – a professional person authorized to practice law; conducts lawsuits or gives legal advice (*non-judgemental reference to person*)

(11) alarm clock, alarm – a clock that wakes sleeper at preset time (*non-judgemental reference to object*)

(12) war, warfare – the waging of armed conflict against an enemy; "thousands of people were killed in the war" (*non-judgemental reference to event*)

3.1.3 Both

In rare cases, a sense can be both subjective and objective (denoted by *B*). The following are the two most frequent cases. First, a WordNet sense might conflate a private state meaning and an objective meaning of a word in the gloss description. Thus, in Example 13 we have the objective literal use of the word *tarnish* mentioned such as *tarnish the silver*, which does not express a private state. However, it also includes a metaphorical use of *tarnish* as in *tarnish a reputation*, which implicitly expresses a negative attitude.

(13) tarnish, stain, maculate, sully, defile—make dirty or spotty, as by exposure to air; also used metaphorically; "The silver was tarnished by the long exposure to the air"; "Her reputation was sullied after the affair with a married man"

The second case includes the inclusion of near-synonyms (Edmonds, 1999) which differs on sentiment in the same synset list. Thus in Example 14, the term *alcoholic* is objective as it is not necessarily judgemental, whereas the other words in the synset such as *soaker* or *souse* are normally insults and therefore subjective.

(14) alcoholic, alky, dipsomaniac, boozer, lush, soaker, souse—a person who drinks alcohol to excess habitually

3.2 Polarity

3.2.1 Polarity of Subjective Senses

The polarity of a subjective sense can be positive (Category *S:P*), negative (*S:N*), or varying, dependent on context or individual preference (*S:V*). The definitions of these three categories are as follows.

- ***S:P*** is assigned to private states that express a positive attitude, emotion or judgement (see Example 7).

- ***S:N*** is assigned to private states that express a negative attitude, emotion or judgement (see Example 5, 6 and 8).

- ***S:V*** is used for senses where the polarity is varying by context or user. For example, it is likely that you give an opinion about somebody if you call him *aloof*; however, only context can determine whether this is positive or negative (see Example 15).

(15) aloof, distant, upstage—remote in manner; “stood apart with aloof dignity”; “a distant smile”; “he was upstage with strangers” (*S:V*)

3.2.2 Polarity of Objective Senses

There are many senses that are objective but have strong negative or positive connotations. For example, *war* describes in many texts an objective state (“He fought in the last war”) but still has strong negative connotations. In many (but not all) cases the negative or positive associations are mentioned in the WordNet gloss. Therefore, we can determine three polarity categories for objective senses:

- ***O:NoPol*** Objective with no strong, generally shared connotations (see Example 9, 10, 11 and 16).
- ***O:P*** Objective senses with strong positive connotations. These refer to senses that do not describe or express a mental state, emotion or judgement but whose presence in a text would give it a strong feel-good flavour (see Example 17).
- ***O:N*** Objective senses with strong negative connotations. These are senses that do not describe or express an emotion or judgement but whose presence in a text would give it a negative flavour (see Example 12). Another example is (18): you can verify objectively whether a liquor was diluted, but it is normally associated negatively.

(16) above—appearing earlier in the same text; “flaws in the above interpretation” (*O:NoPol*)

(17) remedy, curative, cure – a medicine or therapy that cures disease or relieve pain (*O:P*)

(18) adulterate, stretch, dilute, debase—corrupt, debase, or make impure by adding a foreign or inferior substance; often by replacing valuable ingredients with inferior ones; “adulterate liquor” (*O:N*)

We only allow positive and negative annotations for objective senses if we expect *strong* connotations that are *shared* among most people (in Western culture). Thus, for example *war*, *diseases* and *crimes* can relatively safely be predicted to have shared negative connotations. In contrast, a sense like the one of *alarm clock* in Example 11 might have negative connotations for late risers but it would be annotated as *O:NoPol* in our scheme. We are interested in strong shared connotations as the presence of such “loaded” terms can partially indicate bias in a text. In addition, such objective senses are likely to give rise to figurative subjective senses (see Example 18).

4 Experiments and Evaluation

This section describes the experimental setup for our annotation experiments, presents reliability results and discusses the benefits of the use of a hierarchical annotation scheme as well as the problems of bias in the annotation scheme, annotator preferences and bias in the sense inventory.

4.1 Dataset and Annotation Procedure

The dataset used in our annotation scheme is the Micro-WNOp corpus[5], which contains all senses of 298 words in WordNet 2.0. We used it as it is representative of WordNet with respect to its part-of-speech distribution and includes synsets of relatively frequent words, including a wide variety of subjective senses. It contains 1105 synsets in total, divided into three groups *common* (110 synset), *group1* (496 synsets) and *group2* (499 synsets). We used *common* as the training set for the annotators and tested annotation reliability on *group1*.

Annotation was performed by two annotators. Both are fluent English speakers; one is a computational linguist whereas the other is not in linguistics. All annotation was carried out independently and without discussion during the annotation process. The annotators were furnished with guideline annotations with examples for each category. Annotators saw the full synset, including all synonyms, glosses and examples.

4.2 Agreement Study

Training. The two annotators first annotated the *common group* for training. Observed agreement on the training data is 83.6%, with a kappa (Cohen, 1960) of 0.76. Although this looks overall quite good, several categories are hard to identify, for example *B* and *S:V*, as can be seen in the confusion matrix below (Table 1) with Annotator 1 in columns and Annotator 2 in the rows.

Testing. Problem cases were discussed between the annotators and a larger study on *group 1* as test

[5] Available at http://www.unipv.it/wnop/micrownop.tgz

Table 1: Confusion matrix for the training data

	B	S:N	S:P	S:V	O:NoPol	O:N	O:P	total
B	1	0	0	0	2	0	0	3
S:N	0	13	0	0	0	2	0	15
S:P	0	0	8	1	1	0	0	10
S:V	1	1	0	13	6	0	0	21
O:NoPol	1	0	0	0	50	0	0	51
O:N	0	0	0	0	2	4	0	6
O:P	0	0	1	0	0	0	3	4
total	3	14	9	14	61	6	3	110

data was carried out. Table 2 shows the confusion matrix for all 7 categories.

Table 2: Confusion matrix on the test set

	B	S:N	S:P	S:V	O:NoPol	O:N	O:P	total
B	7	2	0	2	0	0	0	11
S:N	0	41	1	0	0	0	0	42
S:P	0	0	65	4	0	0	2	71
S:V	0	0	7	17	3	0	0	27
O:NoPol	9	1	2	6	253	5	8	284
O:N	0	14	0	2	0	25	0	41
O:P	1	0	5	0	1	0	13	20
total	17	58	80	31	257	30	23	496

The observed agreement is 84.9% and the kappa is 0.77. This is good agreement for a relatively subjective task. However, there is no improvement over agreement in training although an additional clarification phase of the training material took place between training and testing.

We also computed single category kappa in order to estimate which categories proved the most difficult. Single category-kappa concentrates on one target category and conflates all other categories into one *non-target* category and measures agreement between the two resulting categories. The results showed that *S:N* (0.80), *S:P* (0.84) and *O:NoPol* (0.86) were highly reliable with less convincing results for *B* (0.49), *S:V* (0.56), *O:N* (0.68), and *O:P* (0.59). *B* is easily missed during annotation (see Example 19), *S:V* is easily confused with several other categories (Example 20), whereas *O:N* is easily confused with *O:NoPol* and *S:N* (Example 21); and *O:P* is easily confused with *O:NoPol* and *S:P* (Example 22).

(19) antic, joke, prank, trick, caper, put-on—a ludicrous or grotesque act done for fun and amusement (*B vs O:NoPol*)

(20) humble—marked by meekness or modesty; not arrogant or prideful; "a humble apology" (*S:V vs S:P*)

(21) hot—recently stolen or smuggled; "hot merchandise"; "a hot car" (*O:N vs O:NoPol*)

(22) profit, gain—the advantageous quality of being beneficial (*S:P vs O:P*)

Our annotation scheme also needs testing on an even larger data set as a few categories such as *B* and *O:P* occur relatively rarely.

4.3 The Effect of Hierarchical Annotation

As mentioned above, our annotation scheme allows us to consider the subjectivity or polarity distinction individually, leaving the full categorization underspecified.

Subjectivity Distinction Only. For subjectivity distinctions we collapse *S:V*, *S:P* and *S:N* into a single label *S* (subjective) and *O:NoPol*, *O:N* and *O:P* into a single label *O* (objective). *B* remains unchanged. The resulting confusion matrix on the test set is in Table 3.

Table 3: Confusion matrix for *Subjectivity*

	B	S	O	total
B	7	4	0	11
S	0	135	5	140
O	10	30	305	345
total	17	169	310	496

Observed agreement is 90.1% and kappa is 0.79. Single category kappa is 0.49 for *B*, 0.82 for *S* and 0.80 for *O*. As *B* is a very rare category (less than 5% of items), this is overall an acceptable level of distinction with excellent reliability for the two main categories.

Polarity Distinction Only. We collapse *O:N* and *S:N* into a single category *N* (negative) and *O:P* and *S:P* into *P* (positive), leaving the other categories intact. This results in 5 categories *B*, *S:V/V*, *NoPol*, *N* and *P*. The resulting confusion matrix is in Table 4.

Table 4: Confusion matrix for *Polarity*

	B	N	P	V	NoPol	total
B	7	2	0	2	0	11
N	0	80	1	2	0	83
P	1	0	85	4	1	91
V	0	0	7	17	3	27
NoPol	9	6	10	6	253	284
total	17	88	103	31	257	496

Observed agreement is 89.1% and kappa is 0.83. Single category kappa is as follows: *B* (0.49), *N* (0.92), *P* (0.85), *V* (0.56), and *NoPol* (0.86). This means all categories but *B* and *V* (together about 10% of items) are reliably identifiable.

Overall we show that both polarity and subjectivity identification of word senses can be reliably annotated and are well-defined tasks for automatic classification. Specifically the percentage agreement of about 90% for word sense polarity/subjectivity identification is substantially higher than the one of 78% reported in Andreeivskaia and Bergler (2006). Agreement for polarity-only is significantly higher than for the full annotation scheme, showing the value of hierarchical annotation. We believe hierarchical annotation is also appropriate for this task, as subjectivity and polarity are linked but still separate concepts. Thus, a researcher might want to mainly focus on explicitly expressed opinions as exemplified by subjectivity, whereas another can also focus on opinion bias in a text as expressed by loaded words of positive or negative polarity.

4.4 Bias in Annotation Performance, Sense Inventory and Annotation Guidelines

Why do annotators assign different labels to some senses? Three main aspects are responsible for non-spurious disagreement.

Firstly, individual perspective or bias played a role. For example, Annotator 2 was more inclined to give positive or negative polarity labels than Annotator 1 as can be seen in Table 4, where Annotator 2 assigned 103 positive and 88 negative labels,whereas Annotator 1 assigned only 91 positive and 83 negative labels.

Secondly, the WordNet sense inventory conflates near-synonyms which just differ in sentiment properties (see Section 3.1.3 and Example 14). Although the labels *B* and *S:V* were specifically created in the annotation scheme to address this problem, these cases still proved confusing to annotators and do not readily lead to consistent annotation.

Thirdly, WordNet sometimes includes a connotation bias either in its glosses or in its hierarchical organization. Here we use the word connotation bias for the inclusion of connotations that seem highly controversial. Thus, in Example 23, the WordNet gloss for *Iran* evokes negative connotations by mentioning allegations of terrorism.[6] In Example 24 *skinhead* is a hyponym of bully, giving strong negative connotations for *all* skinheads. Although the annotation scheme explicitly encourages annotators to disregard especially such controversial connotations as in Example 23 such examples can still confuse annotators and show that word sense annotation is to a certain degree dependent on the sense inventory used.

(23) Iran, Islamic Republic of Iran, Persia—a theocratic islamic republic in the Middle East in western Asia; Iran was the core of the ancient empire that was known as Persia until 1935; rich in oil; involved in state-sponsored terrorism

(24) skinhead ⟵ bully, tough, hooligan, ruffian, roughneck, rowdy, yob, yobo, yobbo

Some of our *good* reliability performance might be due to one particular instance of bias in the annotation guidelines. We strongly advised annotators to only annotate positive or negative polarity for objective senses when strong, shared connotations are expected,[7] thereby "de-individualising" the task of polarity annotation. This introduces a bias towards the category *NoPol* for objective senses. We also did not allow varying polarity for objective senses, instructing annotators that such polarity would be unclear and should be annotated as *NoPol* as not being a strong shared connotation. It can of course be questioned whether the introduction of such a bias is good or not. It helps agreement but might reduce the usefulness of the annotation as individual connotations are not annotated for objective senses. However, to consider more individual connotations needs an annotation effort with a much larger number of annotators to arrive at a profile of polarity connotations over a larger population. We leave this for future work. Our current framework is comprehensive for subjectivity as well as polarity for subjective senses.

4.5 Gold Standard

After discussion between the two annotators, a gold standard annotation was agreed upon. Our data set consists of this agreed set as well as the remainder of the Micro-WNOp corpus (*group2*) annotated by one of the annotators alone after agreement was established.

How many words are subjectivity-ambiguous or polarity-ambiguous, i.e. how much information do we gain by annotating senses over annotating words? As the number of senses increases with word frequency, we expect rare words to be less likely to be subjectivity-ambiguous than frequent words. The Micro-WNOp corpus contains relatively frequent words so we will get an overestimation of subjectivity-ambiguous word types from this corpus, though not necessarily of word tokens. Of all 298 words, 97 (32.5%) are subjectivity-ambiguous, a substantial number. Fewer words are

[6] Note that this was part of WordNet 2.0 and has been removed in WordNet 2.1.

[7] See Section 3.2.2 for justification.

polarity-ambiguous: only 10 words have at least one positive and one negatively annotated sense with a further 44 words having at least one subjective sense with varying polarity (*S:V*). This suggests that subjective and objective uses of the same word are more frequent than reverses in emotional orientation.

5 Comparison to Original Polarity Annotation (Cerini et al.)

We can compare the reliability of our own annotation scheme with the original (polarity) annotation in the Micro-WNOp corpus. Cerini et al. (2007) do not present agreement figures but as their corpus is publically available we can easily compute reliability. Recall that each synset has a triplet of numerical scores between 0 and 1 each: positivity, negativity and neutrality, which is not explicitly annotated but derived as $1 - (positivity + negativity)$. Subjectivity in our sense (existence of a private state) is *not* annotated.

The ratings of three annotators are available for *Group 1* and of two annotators for *Group 2*. We measured the Pearson correlation coefficient between each annotator pair for both groups for both negativity and positivity scoring. As correlation can be high without necessarily high agreement on absolute values, we also computed a variant of kappa useful for numerical ratings, namely alpha (Artstein and Poesio, 2005), which gives weight to degrees of disagreement. Thus, a disagreement between two scores would be weighted as the absolute value of $score1 - score2$. The results are listed in Table 5.

Table 5: Reliability of original annotation on Micro-WNOp

dataset	raters	score type	correlation	alpha
Group 1	1 and 2	negative	83.7	64.9
Group 1	1 and 3	negative	86.4	71.8
Group 1	2 and 3	negative	82.5	56.9
Group 1	1 and 2	positive	80.5	60.9
Group 1	1 and 3	positive	87.8	74.9
Group 1	2 and 3	positive	78.2	57.5
Group 2	1 and 2	negative	95.9	90.7
Group 2	1 and 2	positive	92.2	84.9

Correlation between the annotators is high. However, Rater 2 (in Group1) still behaves differently from the other two raters, giving consistently higher or lower scores overall, leading to low alpha. Thus, we can conclude that Group 2 is much more reliably annotated than Group 1 and that especially Rater 2 in Group 1 is an outlier in this (small) set of raters. This also shows that work with several annotators is valuable and should be conducted for our scheme as well.

6 Conclusion and Future Work

We elicit human judgements on the subjectivity and polarity of word senses. To the best of our knowledge, this is the first such annotation scheme for both categories. We detail the definitions for each category and measure the reliability of the annotation. The experimental results show that when using all 7 categories, only 3 categories (*S:N*, *S:P*, and *O:NoPol*) are reliable while the reliability of the other 4 categories is not high. We also show that this is improved by the virtue of hierarchical annotation and that the general tasks of subjectivity and polarity annotation on word senses are therefore well-defined. Moreover, we also discuss the effect of different kinds of bias on our approach.

In future we will refine the guidelines for the more difficult categories, including more detailed advice on how to deal with sense inventory bias. We will also perform larger-scale annotation exercises with more annotators as the latter is necessary to deal with more individualised polarity connotations. In addition, we will use the data to test learning methods for the automatic detection of subjectivity and polarity properties of word senses.

References

Andreevskaia, Alina and Sabine Bergler. 2006. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. *Proceedings of EACL'06.*

Artstein, Ron and Massimo Poesio. 2005. $Kappa^3$=alpha(or beta). *Technical Report CSM-437, University of Essex.*

Bradley, Margaret and Peter Lang. 1999. Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings *Technical report C-1, the Center for Research in Psychophysiology, University of Florida. .*

Cerini, Sabrina, Valentina Compagnoni, Alice Demontis, Maicol Formentelli, and Caterina Gandini. 2007. Micro-WNOp: A Gold Standard for the Evaluation of Automatically Compiled Lexical Resources for Opinion Mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics.*

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, Vol.20, No.1.*

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. *Proceedings of APFA'01.*

Edmonds, Philip. 1999. Semantic Representations Of Near-Synonyms For Automatic Lexical Choice. *PhD thesis, University of Toronto.*

Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC'06.*

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of ACL'97.*

Heise, David. 2001. Project Magellan: Collecting Cross-culture Affective Meanings via the Internet. Electronic Journal of Sociology.

Osgood, Charles, William May, and Murray Miron. 1975. Cross-cultural Universals of Affective Meaning. *University of Illinois Press.*

Osgood, Charles, George Suci, and Percy Tannenbaum. 1957. The Measurment of Meaning. *University of Illinois Press.*

Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet-Affect: an Affective Extension of WordNet. *Proceedings of LREC'04.*

Takamura, Hiroya, Takashi Inui, and Manabu Okumura. 2005. Extracting Semantic Orientations of Words using Spin Model. *Proceedings of ACL'05.*

Wiebe, Janyce and Rada Micalcea. 2006. Word Sense and Subjectivity. *Proceedings of ACL'06.*

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation.*

Human Judgements in Parallel Treebank Alignment

Martin Volk and Torsten Marek
University of Zurich
Institute of Computational Linguistics
8050 Zurich, Switzerland
volk@cl.uzh.ch

Yvonne Samuelsson
Stockholm University
Department of Linguistics
106 91 Stockholm, Sweden
yvonne.samuelsson@ling.su.se

Abstract

We have built a parallel treebank that includes word and phrase alignment. The alignment information was manually checked using a graphical tool that allows the annotator to view a pair of trees from parallel sentences. We found the compilation of clear alignment guidelines to be a difficult task. However, experiments with a group of students have shown that we are on the right track with up to 89% overlap between the student annotation and our own. At the same time these experiments have helped us to pin-point the weaknesses in the guidelines, many of which concerned unclear rules related to differences in grammatical forms between the languages.

1 Introduction

Establishing translation correspondences is a difficult task. This task is traditionally called alignment and is usually performed on the paragraph level, sentence level and word level. Alignment answers the question: Which part of a text in language L1 corresponds in meaning to which part of a text in language L2 (under the assumption that the two texts represent the same meaning in different languages). This may mean that one text is the translation of the other or that both are translations derived from a third text.

There is considerable interest in automating the alignment process. Automatic sentence alignment of legacy translations helps to fill translation memories. Automatic word alignment is a crucial step in training statistical machine translation systems. Both sentence and word alignment have to deal with 1:many alignments, i.e. sometimes a sentence in one language is translated as two or three sentences in the other language.

In other respects sentence alignment and word alignment are fundamentally different. It is relatively safe to assume the same sentence order in both languages when computing sentence alignment. But such a monotonicity assumption is not possible for word alignment which needs to allow for word order differences and thus for crossing alignments. And while algorithms for sentence alignment usually focus on length comparisons (in terms of numbers of characters), word alignment algorithms use cross-language cooccurrence frequencies as a key feature.

Our work focuses on word alignment and on an intermediate alignment level which we call phrase alignment. Phrase alignment encompasses the alignment from simple noun phrases and prepositional phrases all the way to complex clauses. For example, on the word alignment level we want to establish the correspondence of the German "verb form plus separated prefix" *fing an* with the English verb form *began*. While in phrase alignment we mark the correspondence of the verb phrases *ihn in den Briefkasten gesteckt* and *dropped it in the mail box*.

We regard phrase alignment as alignment between linguistically motivated phrases, in contrast to some work in statistical machine translation where phrase alignment is defined as the alignment between arbitrary word sequences. Our phrase alignment is alignment between nodes in constituent structure trees. See figure 1 for an example of a tree pair with word and phrase alignment.

We believe that such linguistically motivated phrase alignment provides useful phrase pairs for example-based machine translation, and provides interesting insights for translation science and cross-language comparisons. Phrase alignments are particularly useful for annotating correspondences of idiomatic or metaphoric language use.

2 The Parallel Treebank

We have built a trilingual parallel treebank in English, German and Swedish. The treebank consists of around 500 trees from the novel Sophie's World and 500 trees from economy texts (an annual report from a bank, a quarterly report from an international engineering company, and the banana certification program of the Rainforest Alliance). The sentences in Sophie's World are relatively short (14.8 tokens on average in the English version), while the sentences in the economy texts are much longer (24.3 tokens on average; 5 sentences in the English version have more than 100 tokens).

The treebanks in English and German consist of constituent structure trees that follow the guidelines of existing treebanks, the NEGRA/TIGER guidelines for German and the Penn treebank guidelines for English. There were no guidelines for Swedish constituent structure trees. We have therefore adapted the German treebank guidelines for Swedish. Both German trees and Swedish trees are annotated with flat structures but subsequently automatically deepened to result in richer and linguistically more plausible tree structures.

When the monolingual treebanks were finished, we started with the word and phrase alignment. For this purpose we have developed a special tool called the Stockholm TreeAligner (Lundborg et al., 2007) which displays two trees and allows the user to draw alignment lines by clicking on nodes and words. This tool is similar to word alignment tools like ILink (Ahrenberg et al., 2003) or Cairo (Smith and Jahr, 2000). As far as we know our tool is unique in that it allows the alignments of linguistically motivated phrases via node alignments in parallel constituent structure trees (cf. (Samuelsson and Volk, 2007)).

After having solved the technical issues, the challenge was to compile precise and comprehensive guidelines to ensure smooth and consistent alignment decisions. In (Samuelsson and Volk, 2006) we have reported on a first experiment to evaluate inter-annotator agreement from our alignment tasks.

In this paper we report on another recently conducted experiment in which we tried to identify the weaknesses in our alignment guidelines. We asked 12 students to alignment 20 tree pairs (English and German) taken from our parallel treebank. By comparing their alignments to our Gold Standard and to each other we gained valuable insights into the difficulty of the alignment task and the quality of our guidelines.

3 Related Research

Our research on word and phrase alignment is related to previous work on word alignment as e.g. in the Blinker project (Melamed, 1998) or in the UPLUG project (Ahrenberg et al., 2003). Alignment work on parallel treebanks is rare. Most notably there is the Prague Czech-English treebank (Kruijff-Korbayová et al., 2006) and the Linköping Swedish-English treebank (Ahrenberg, 2007). There has not been much work on the alignment of linguistically motivated phrases. Tinsley et al. (2007) and Groves et al. (2004) report on semi-automatic phrase alignment as part of their research on example-based machine translation.

Considering the fact that the alignment task is essentially a semantic annotation task, we may also compare our results to other tasks in semantic corpus annotation. For example, we may consider the methods for resolving annotation conflicts and the figures for inter-annotator agreement in frame-semantic annotation as found in the German SALSA project (cf. (Burchardt et al., 2006)).

4 Our Alignment Guidelines

We have compiled alignment guidelines for word and phrase alignment between annotated syntax trees. The guidelines consist of general principles, concrete rules and guiding principles.

The most important general principles are:

1. Align items that can be re-used as units in a machine translation system.

2. Align as many items (i.e. words and phrases) as possible.

3. Align as close as possible to the tokens.

The first principle is central to our work. It defines the general perspective for our alignment.

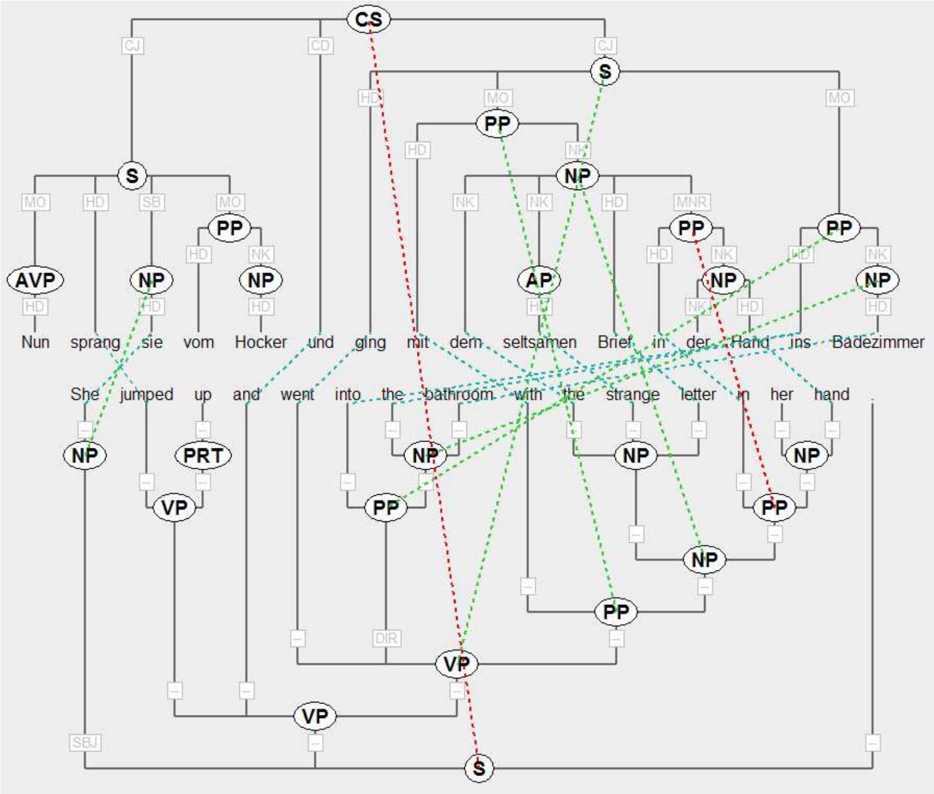

Figure 1: Tree pair German-English with word and phrase alignments.

We do not want to know which part of a sentence has possibly given rise to which part of the correspondence sentence. Instead our perspective is on whether a phrase pair is general enough to be re-used as translation unit in a machine translation system. For example, we do not want to align *die Verwunderung über das Leben* with *their astonishment at the world* although these two phrases were certainly triggered by the same phrase in the original and both have a similar function in the two corresponding sentences. These two phrases seen in isolation are too far apart in meaning to license their re-use. We are looking for correspondences like *was für eine seltsame Welt* and *what an extraordinary world* which would make for a good translation in many other contexts.

Some special rules follow from this principle. For example, we have decided that a pronoun in one language shall never be aligned with a full noun in the other, since such a pair is not directly useful in a machine translation system.

Principles 2 and 3 are more technical. Principle 2 tells our annotators that alignment should be exhaustive. We want to re-use as much as possible from the treebank, so we have to look for as many alignments as possible. And principle 3 says that in case of doubt the alignment should go to the node that is closest to the terminals. For example, our German treebank guidelines require a multi-word proper noun to first be grouped in a PN phrase which is a daughter node of a noun phrase `[[Sofie Amundsen]PN ]NP` whereas the English guidelines only require the NP node `[Sophie Amundsen]NP`. When we align the two names, principle 3 tells us to draw the alignment line between the German PN node and the English NP node since the PN node is closer to the tokens than the German NP node.

Often we are confronted with phrases that are not exact translation correspondences but approximate translation correspondences. Consider the phrases *mehr als eine Maschine* and *more than a piece of hardware*. This pair does not represent the closest possible translation but it represents a possible translation in many contexts. In a way we could classify this pair as the “second-best” translation. To allow for such distinctions we provide our annotators with a choice between exact translation correspondences and approximate correspondences. We also use the term **fuzzy correspondence** to refer to and give an intuitive picture of these approximate correspondences. The option to distinguish between different alignment strengths sounded very attractive at the start but it turned out to be the source for some headaches later. Where and how can we draw the line between exact and fuzzy translation correspondences?

We have formulated some clear-cut rules:

1. If an acronym is to be aligned with a spelled-out term, it is always an approximate alignment. For example, in our economy reports the English acronym *PT* stands for *Power Technology* and is aligned to the German *Energietechnik* as a fuzzy correspondence.

2. Proper names shall be aligned as exact alignments (even if they are spelled differently across languages; e.g. *Sofie* vs. *Sophie*).

But many open questions persist. Is *einer der ersten Tage im Mai* an exact or rather a fuzzy translation correspondence of *early May*? We decided that it is not an exact correspondence. How shall we handle *zu dieser Jahreszeit* vs. *at this time of the year* where a literal translation would be *in this season*? We decided that the former is still an exact correspondence. These examples illustrate the difficulties that make us wonder how useful the distinction between exact and approximate translation correspondence really is.

Automatically ensuring the overall consistency of the alignment decisions is a difficult task. But we have used a tool to ensure the consistency within the exact and approximate alignment classes. The tool computes the token span for each alignment and checks if the same tokens pairs have always received the same alignment type. For example, if the phrase pair *mit einer blitzschnellen Bewegung* and *with a lightning movement* is once annotated as exact alignment, then it should always be annotated as exact alignment. Figure 1 shows approximate alignments between the PPs *in der Hand* and *in her hand*. It was classified as approximate rather than exact alignment since the German PP lacks the possessive determiner.

Currently our alignment guidelines are 6 pages long with examples for English-German and English-Swedish alignments.

5 Experiments with Student Annotators

In order to check the inter-annotator agreement for the alignment task we performed the following experiment. We gave 20 tree pairs in German and English to 12 advanced undergraduate students in a class on ”Machine Translation and Parallel Corpora”. Half of the tree pairs were taken from our Sophie’s World treebank and the other half from our Economy treebank. We made sure that there was one 1:2 sentence alignment in the sample. The students did not have access to the Gold Standard alignment.

In class we demonstrated the alignment tool to the students and we introduced the general alignment principles to them. Then the students were given a copy of the alignment guidelines. We asked them to do the alignments independently of each other and to the best of their knowledge according to the guidelines.

In our own annotation of the 20 tree pairs (= the Gold Standard alignment) we have the following numbers of alignments:

	type	exact	fuzzy	total
Sophie part	word	75	3	78
	phrase	46	12	58
Economy part	word	159	19	178
	phrase	62	9	71

In the Sophie part of the experiment treebank we have 78 word-to-word alignments and 58 phrase-to-phrase alignments. Note that some phrases consist only of one word and thus the same alignment information is represented twice. We have deliberately kept this redundancy.

The alignments in the Sophie part consist of 125 times 1:1 alignments, 4 times 1:2 alignments and one 1:3 alignment (*wäre* vs. *would have been*) when viewed from the German side. There are 3 times 1:2 alignments (e.g. *introducing* vs. *stellte vor*) and no other 1:many alignment when viewed from the English side.

In the Economy part the picture is similar. The vast majority are 1:1 alignments. There are 207 times 1:1 alignments and 21 times 1:2 alignments (many of which are German compound nouns) when viewed from German. And there are 235 times 1:1 alignments, plus 4 times 1:2 alignments, plus 2 times 1:3 alignments when viewed from English (e.g. the *Americas* was aligned to the three tokens *Nord- und Südamerika*).

The student alignments showed a huge variety in terms of numbers of alignments. In the Sophie part they ranged from 125 alignments to bare 47 alignments (exact alignments and fuzzy alignments taken together). In the Economy part the variation was between 259 and 62 alignments.

On closer inspection we found that the student with the lowest numbers works as a translator and chose to use a very strict criterion of translation equivalence rather than translation correspondence. Three other students at the end of the list are not native speakers of either German and English. We therefore decided to exclude these 4 students from the following comparison.

The student alignments allow for the investigation of a number of interesting questions:

1. How did the students' alignments differ from the Gold Standard?
2. Which were the alignments done by all students?
3. Which were the alignments done by single students only?
4. Which alignments varied most between exact and fuzzy alignment?

When we compared each student's alignments to the Gold Standard alignments, we computed three figures:

1. How often did the student alignment and the Gold Standard alignment overlap?
2. How many Gold Standard alignments did the student miss?
3. How many student alignments were not in the Gold Standard?

The remaining 8 students reached between 81% and 48% overlap with the Gold Standard on the Sophie part, and between 89% and 66% overlap with the Gold Standard on the Economy texts. This can be regarded as their recall values if we assume that the Gold Standard represents the correct alignments. These same 8 students additionally had between 2 and 22 own alignments in the Sophie part and between 12 and 55 own alignments in the Economy part.

So the interesting question is: What kind of alignments have they missed, and which were the additional own alignments that they suggested (alignments that are not in the gold standard)? We first checked the students with the highest numbers of own alignments. We found that some of these alignments were due to the fact that students had ignored the rule to align as close to the tokens as possible (principle 3 above).

Another reason was that students sometimes aligned a word (or some words) with a node. For example, one student had aligned the word *natürlich* to the phrase *of course* instead of to the word sequence *of course*. Our alignment tool allows that, but the alignment guidelines discourage such alignments. There might be exceptional cases where a word-to-phrase alignment is necessary in order to keep valuable information, but in general we try to stick to word-to-word and phrase-to-phrase alignments.

Another discrepancy occurred when the students aligned a German verb group with a single verb form in English (e.g. *ist zurückzuführen* vs. *reflecting*). We have decided to only align the full verb to the full verb (independent of the inflection). This means that we align only *zurückzuführen* to *reflecting* in this example.

The uncertainties on how to deal with different grammatical forms led to the most discrepancies. Shall we align the definite NP *die Umsätze* with the indefinite NP *revenues* since it is much more common to drop the article in an English plural NP than in German? Shall we align a German genitive NP with an of-PP in English (*der beiden Divisionen* vs. *of the two divisions*)? We have decided to give priority to form over function and thus to align the NP *der beiden Divisionen* with the NP *the two divisions*. But of course this choice is debatable.

When we compute the intersection of the alignments done by all students (ignoring the difference between exact and fuzzy alignments), we find that about 50% of the alignments done by the student with the smallest number of alignments is shared by all other students. All of the alignments in the intersection are in our Gold Standard file. This indicates that there is a core of alignments that are obvious and uncontroversial. Most of them are word alignments.

When we compute the union of the alignments done by all students (again ignoring the difference between exact and fuzzy alignments), we find that the number of alignments in the union is 40% to 50% higher than the number of alignments done by the student with the highest number of alignments. It is also about 40% to 50% higher than the number of alignments in the Gold Standard. This means that there is considerable deviation from the Gold Standard.

Comparing the union of the students' alignments to the Gold Standard points to some weaknesses of the guidelines. For example, one alignment in the Gold Standard that was missed by all students concerns the alignment of a German pronoun (*wenn* ***sie*** *die Hand ausstreckte*) to an empty token in English (*herself* __ *shaking hands*). Our guidelines recommend to align such cases as fuzzy alignments, but of course it is difficult to determine that the empty token really corresponds to the German word.

Other discrepancies concern cases of differing grammatical forms, e.g. a German definite singular noun phrase (*die Hand*) that was aligned to an English plural noun phrase (*Hands*) in the Gold Standard but missed by all students. Finally there are a few cases where obvious noun phrase correspondences were simply overlooked by all students (*sich - herself*) although the tokens themselves were aligned. Such cases should be handled by an automated process in the alignment tool that projects from aligned tokens to their mother nodes (in particular in cases of single token phrases).

We also investigated how many exact alignments and how many fuzzy alignments the students had used. The following table gives the figures.

	exact	fuzzy	overlap	total
Sophie part	152	106	69	189
Economy part	296	188	119	366

The alignments done by all students resulted in a union set of 189 alignments for the Sophie part and 366 alignments for the Economy part. The alignments in the Sophie part consisted of 152 exact alignments and 106 fuzzy alignments. This means that 69 alignments were marked as both exact and fuzzy. In other words, in 69 cases at least one student has marked an alignment as fuzzy while at least one other student has marked the same alignment as good. So there is still considerable confusion amongst the annotators on how to decide between exact and fuzzy alignments. And in case of doubt many students have decided in favor of fuzzy alignments.

6 Conclusions

We have shown the difficulties in creating cross-language word and phrase alignments. Experiments with a group of students have helped to identify the weaknesses in our alignment guidelines and in our Gold Standard alignment. We have realized that the guidelines need to contain a host of fine-grained alignment rules and examples that will clarify critical cases.

In order to evaluate a set of alignment experiments with groups of annotators it is important to have good visualization tools to present the results. We have worked with Perl scripts for the comparison and with our own TreeAligner tool for the visualization. For example we have used two colors to visualize a student's alignment overlap with the Gold Standard in one color and his own alignments (that are not in the Gold Standard) in another color.

In order to visualize the agreements of the whole group it would be desirable to have the option to increase the alignment line width in proportion to the number of annotators that have chosen a particular alignment link. This would give an intuitive impression of strong alignment links and weak alignment links.

Another option for future extension of this work is an even more elaborate classification of the alignment links. (Hansen-Schirra et al., 2006) have demonstrated how a fine-grained distinction between different alignment types could look like. Annotating such a corpus will be labor-intensive but provide for a wealth of cross-language observations.

References

Ahrenberg, Lars, Magnus Merkel, and Michael Petterstedt. 2003. Interactive word alignment for language engineering. In *Proc. of EACL-2003*, Budapest.

Ahrenberg, Lars. 2007. LinES: An English-Swedish parallel treebank. In *Proc. of Nodalida*, Tartu.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974, Genoa.

Groves, Declan, Mary Hearne, and Andy Way. 2004. Robust sub-sentential alignment of phrase-structure trees. In *Proceedings of Coling 2004*, pages 1072–1078, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Hansen-Schirra, Silvia, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proceedings of the EACL Workshop on Multidimensional Markup in Natural Language Processing (NLPXML-2006)*, pages 35– 42, Trento.

Kruijff-Korbayová, Ivana, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for the Czech-English word alignment. In *Proceedings of LREC*, Genova.

Lundborg, Joakim, Torsten Marek, Maël Mettler, and Martin Volk. 2007. Using the Stockholm TreeAligner. In *Proc. of The 6th Workshop on Treebanks and Linguistic Theories*, Bergen, December.

Melamed, Dan. 1998. Manual annotation of translational equivalence: The blinker project. Technical Report 98-06, IRCS, Philadelphia PA.

Samuelsson, Yvonne and Martin Volk. 2006. Phrase alignment in parallel treebanks. In Hajic, Jan and Joakim Nivre, editors, *Proc. of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 91–102, Prague, December.

Samuelsson, Yvonne and Martin Volk. 2007. Alignment tools for parallel treebanks. In *Proceedings of GLDV Frühjahrstagung 2007*.

Smith, Noah A. and Michael E. Jahr. 2000. Cairo: An alignment visualization tool. In *Proc. of LREC-2000*, Athens.

Tinsley, John, Ventsislav Zhechev, Mary Hearne, and Andy Way. 2007. Robust language pair-independent sub-tree alignment. In *Machine Translation Summit XI Proceedings*, Copenhagen.

An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text

Plaban Kr. Bhowmick, Pabitra Mitra, Anupam Basu
Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India – 721302
{plaban,pabitra,anupam}@cse.iitkgp.ernet.in

Abstract

An affective text may be judged to belong to multiple affect categories as it may evoke different affects with varying degree of intensity. For affect classification of text, it is often required to annotate text corpus with affect categories. This task is often performed by a number of human judges. This paper presents a new agreement measure inspired by Kappa coefficient to compute inter-annotator reliability when the annotators have freedom to categorize a text into more than one class. The extended reliability coefficient has been applied to measure the quality of an affective text corpus. An analysis of the factors that influence corpus quality has been provided.

1 Introduction

The accuracy of a supervised machine learning task primarily depends on the annotation quality of the data, that is used for training and cross validation. Reliability of annotation is a key requirement for the usability of an annotated corpus. Inconsistency or noisy annotation may lead to the degradation of performances of supervised learning algorithms. The data annotated by a single annotator may be prone to error and hence an unreliable one. This also holds for annotating an affective corpus, which is highly dependent on the mental state of the subject. The recent trend in corpus development in NLP is to annotate corpus by more than one annotators independently. In corpus statistics, the corpus reliability is measured by coefficient of agreement. The coefficients of agreement are applied to corpus for various goals like measuring *reliability, validity* and *stability* of corpus (Artstein and Poesio, 2008).

Jacob Cohen (Cohen, 1960) introduced Kappa statistics as a coefficient of agreement for nominal scales. The Kappa coefficient measures the proportion of observed agreement over the agreement by chance and the maximum agreement attainable over chance agreement considering pairwise agreement. Later Fleiss (Fleiss, 1981) proposed an extension to measure agreement in ordinal scale data.

Cohen's Kappa has been widely used in various research areas. Because of its simplicity and robustness, it has become a popular approach for agreement measurement in the area of electronics (Jung, 2003), geographical informatics (Hagen, 2003), medical (Hripcsak and Heitjan, 2002), and many more domains.

There are other variants of Kappa like agreement measures (Carletta, 1996). Scott's π (Scott, 1955) was introduced to measure agreement in survey research. Kappa and π measures differ in the way they determine the chance related agreements. π-like coefficients determine the chance agreement among arbitrary coders, while κ-like coefficients treats the chance of agreement among the coders who produced the reliability data (Artstein and Poesio, 2008).

One of the drawbacks of π and Kappa like coefficients except Fleiss' Kappa (Fleiss, 1981) is that they treat all kinds of disagreements in the same manner. Krippendorff's α (Krippendorff, 1980) is a reliability measure which treats different kind of disagreements separately by introducing a notion of distance between two categories. It offers a way

Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, pages 58–65
Manchester, August 2008

to measure agreement in nominal, interval, ordinal and ratio scale data.

Reliability assessment of corpus is an important issue in corpus driven natural language processing and the existing reliability measures have been used in various corpus development tasks. For example, Kappa coefficient has been used in developing parts of speech corpus (Mieskes and Strube, 2006), dialogue act tagging efforts like MapTask (Carletta et al., 1997) and Switchboard (Stolke et al., 1997), subjectivity tagging task (Bruce and Wiebe, 1999) and many more.

The π and κ coefficients measure the reliability of the annotation task where a data item can be annotated with one category. (Rosenberg and Binkowski, 2004) puts an effort towards measuring corpus reliability for multiply labeled data points. In this measure, the annotators are allowed to mark one data point with at most two classes, one of which is primary and other is secondary. This measure was used to determine the reliability of a email corpus where emails are assigned with primary and secondary labels from a set of email types.

Affect recognition from text is a recent and promising subarea of natural language processing. The task is to classify text segments into appropriate affect categories. The supervised machine learning techniques, which requires a reliable annotated corpus, may be applied for solving the problem. In general, a blend of emotions is common in both verbal and non-verbal communication. Unlike conventional annotation tasks like POS corpus development, where one data item may belong to only one category, in affective text corpus, a data item may be fuzzy and may belong to multiple affect categories. For example, the following sentence may belong to *disgust* and *sad* category since it may evoke both the emotions to different degrees of intensity.

> *A young married woman was burnt to death allegedly by her in-laws for dowry.*

This property makes the existing agreement measures inapplicable for determining agreement in emotional corpus. Craggs and Wood (2004) adopted a categorical scheme for annotating emotion in affective text dialogue. They claimed to address the problem of agreement measurement for the data set where one data item may belong to more than one category using an extension of Krippendorff's α. But the details of the extension is yet to be disseminated.

In this paper, we propose a new agreement measure for multiclass annotation which we denote by A_m. The new measure is then applied to an affective text corpus to

- *Assess Reliability:* To test whether the corpus can be used for developing computational affect recognizer.

- *Determine Gold Standard:* To define a gold standard that will be used to test the accuracy of the affect recognizer.

In section 2, we describe the affective text corpus and the annotation scheme. In section 3, we propose a new reliability measure (A_m) for multiclass annotated data. In section 4, we provide an algorithm to determine *gold standard* data from the annotation and in section 5, we discuss about applying A_m measure to the corpus developed by us and some observations related to the annotation.

2 Affective Text Corpus and Annotation Scheme

The affective text corpus collected by us consists of 1000 sentences extracted from *Times of India* news archive[1]. The sentences were collected from headlines as well as articles belonging to political, social, sports and entertainment domain.

Selection of affect categories is a very crucial and important decision problem due to the following reasons.

- The affect categories should be applicable to the considered genre.

- The affect categories should be identifiable from language.

- The categories should be unambiguous.

We shall try to validate these points based on the results obtained, after applying the our extended measure on the text corpus with respect to a set of selected basic emotional categories.

Basic emotions are those for which the respective expressions across culture, ethnicity, age, sex, social structure are invariant (Ortony and Turner, 1990). But unfortunately, there is a long persistent debate among the psychologists regarding

[1] http://timesofindia.indiatimes.com/archive.cms

the number of basic emotional categories (Ortony and Turner, 1990). One of the theories behind the basic emotions is that they are biologically primitive because they possess evolutionary significance related to the basic needs for the survival of the species (Plutchik, 1980). The universality of recognition of emotions from distinctive facial expressions is an indirect technique to establish the basic emotions (Darwin, 1965).

Six basic affect categories (Ekman, Friesen and Ellsworth, 1982) have been considered in emotion recognition from speech (Song et al., 2004), facial expression (Pantic and Rothkrantz, 2000). Our annotation scheme considers six basic emotions, namely, *Anger, Disgust, Fear, Happiness, Sadness, Surprise* as specified by Ekman for affect recognition in text corpus.

The annotation scheme considers the following points:

- Two types are sentences are collected for annotation.
    - *Direct Affective Sentence:* Here, the agent present in the sentence is experiencing a set of emotions, which are explicit in the sentence. For example, in the following sentence *Indian supporters* are the agents experiencing a disgust emotion.

      > *Indian supporters are disgusted about players' performances in the World Cup.*

    - *Indirect Affective Sentence:* Here, the reader of the sentence is experiencing a set of emotions. In the following sentence, the reader is experiencing a *disgust* emotion because the event of *accepting bribe*, is an indecent act carried out by responsible agents like *Top officials*.

      > *Top officials are held for accepting bribe from a poor villager.*

- A sentence may trigger multiple emotions simultaneously. So, one annotator may classify a sentence to more than one affective categories.

- For each emotion, the keywords that trigger the particular emotion are marked.

- For each emotion, the events or objects that trigger the concerned emotion are marked.

Here, we aim at measuring the agreement in annotation. The focus is to measure the agreement in annotation pattern rather than the agreement in individual emotional classes.

3 Proposed Agreement Measure

To overcome the shortcomings of existing reliability measures mentioned earlier, we propose A_m measure, which is an agreement measure for corpus annotation task considering multiclass classification. We present the notion of agreement below.

3.1 Notion of Paired Agreement

In order to allow for multiple labels, we calculate agreement between all the pairs of possible labels. Let *C1* and *C2* be two affect categories, e.g., *anger* and *disgust*. Let <*C1, C2*> denote the category pair. An annotator's assignment of labels can be represented as a pair of binary choices for each category pair <*C1, C2*>, namely, $< 0, 0 >$, $< 0, 1 >$, $< 1, 0 >$, and $< 1, 1 >$. It should be noted that the proposed metric considers the non-inclusion in a category by an annotator pair as an agreement.

For an item, two annotators *U1* and *U2* are said to agree on <*C1, C2*> if the following conditions hold.

$$U1.C1 = U2.C1$$
$$U1.C2 = U2.C2$$

where $U_i.C_j$ signifies that the value for C_j for annotator U_i and the value may either be 1 or 0. For example, if one coder marks an item with *anger* and another with *disgust*, they would disagree on the pairs that include these labels, but still agree that the item does not express *happiness* and *sadness*.

3.2 A_m Agreement Measure

With the notion of paired agreement discussed earlier, the *observed agreement*(P_o) is the proportion of items the annotators agreed on the category pairs and the *expected agreement*(P_e) is the proportion of items for which agreement is expected by chance when the items are randomly. Following the line of Cohen's Kappa (Cohen, 1960), A_m is defined as the proportion of agreement after expected or chance agreement is removed from consideration and is given by

$$A_m = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

When P_o equals P_e, A_m value is computed to be 0, which signifies no non-random agreement among the annotators. An A_m value of 1, the upper limit of A_m, indicates a perfect agreement among the annotators. We define P_o and P_e as follows.

Observed Agreement (P_o)**:**
Let **I** be the number of items, **C** is the number of categories and **U** is the number of annotators and **S** be the set of all category pairs with cardinality $\binom{\mathbf{C}}{2}$. The total agreement on a category pair p for an item i is n_{ip}, the number of annotator pairs who agree on p for i.

The average agreement on a category pair p for an item i is n_{ip} divided by the total number of annotator pairs and is given by

$$P_{ip} = \frac{1}{\binom{\mathbf{U}}{2}} n_{ip} \tag{2}$$

The average agreement for the item i is the mean of P_{ip} over all category pairs and is given by

$$P_i = \frac{1}{\binom{\mathbf{C}}{2}\binom{\mathbf{U}}{2}} \sum_{p \in \mathbf{S}} n_{ip} \tag{3}$$

The observed agreement is the average agreement over all the item and is given by

$$\begin{aligned} P_o &= \frac{1}{\mathbf{I}} \sum_{i=1}^{I} P_i \\ &= \frac{1}{\mathbf{I}\binom{\mathbf{C}}{2}\binom{\mathbf{U}}{2}} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip} \qquad (4) \\ &= \frac{4}{\mathbf{IC}(\mathbf{C}-1)\mathbf{U}(\mathbf{U}-1)} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip} \end{aligned}$$

Expected Agreement (P_e**):**
The expected agreement is defined as the agreement among the annotators when they assign the items to a set of categories randomly. However, since we are considering the agreement on category pairs, we consider the expected agreement to be the expectation that the annotators agree on a category pair. For a category pair, four possible assignment combinations constitute a set which is given by

$$G = \{[0\ 0], [0\ 1], [1\ 1]\}.$$

It is to be noted that the combinations [0 1] and [1 0] are clubbed to one element as they are symmetric to each other. Let $\hat{P}(p_g|u)$ be the overall proportion of items assigned with assignment combination $g \in G$ to category pair $p \in S$ by annotator u and $n_{p_g u}$ be the total number of assignments of items by annotator u with assignment combination g to category pair p. Then $\hat{P}(p_g|u)$ is given by

$$\hat{P}(p_g|u) = \frac{n_{p_g u}}{\mathbf{I}} \tag{5}$$

For an item, the probability that two arbitrary coders agree with the same assignment combination in a category pair is the joint probability of individual coders making this assignments independently. For two annotators u_x and u_y the joint probability is given by $\hat{P}(p_g|u_x)\hat{P}(pg|u_y)$. The probability that two arbitrary annotators agree on a category pair p with assignment combination g is the average over all annotator pairs belonging to W, the set of annotator pairs and is given by

$$\hat{P}(p_g) = \frac{1}{\binom{\mathbf{U}}{2}} \sum_{(u_x,u_y) \in W} \hat{P}(p_g|u_x)\hat{P}(p_g|u_y) \tag{6}$$

The probability that two arbitrary annotators agree on a category pair for all assignment combinations is given by

$$\hat{P}(p) = \sum_{pg \in G} \hat{P}(p_g) \tag{7}$$

The chance agreement is calculated by taking average over all category pairs.

$$P_e = \frac{1}{\binom{\mathbf{C}}{2}} \sum_{p \in S} \hat{P}(p) \tag{8}$$

The A_m measure may be calculated based on the expressions of P_o and P_e as given in Equation 4 and Equation 8 to compute the reliability of annotation with respect to multiclass annotation.

4 Gold Standard Determination

Gold standard data is used as a reference data set for various goals like

- Building reliable classifier

- Determine the performance of a classifier

To attach a set of labels to a data item in the gold standard data, we assign the majority decision label to an item. Let n_O be the number of annotators, who have assigned an item i into category C and n_ϕ annotators have decided not to assign the same item into that category. Then i is assigned to C if $n_O > n_\phi$; otherwise it is not assigned to that category.

Algorithm 1: Algorithm for determining gold standard data

```
Input: Set of I items annotated into C
       categories by U annotators
Output: Gold standard data
foreach annotator $u \in U$ do
    $\xi_u \leftarrow 0$;
end
foreach item $i \in I$ do
    foreach category $c \in C$ do
        $\Theta$ = set of annotators who have
        assigned $i$ in category $c$;
        $\phi$ = set of annotators who have not
        assigned $i$ in category $c$;
        if cardinality($\Theta$)>cardinality($\phi$) then
            assign label $c$ to $i$;
            $\xi_j \leftarrow \xi_j + 1$ where $j \in \Theta$;
        end
        else if cardinality($\Theta$)<cardinality($\phi$)
        then
            do not assign label $c$ to $i$;
            $\xi_j \leftarrow \xi_j + 1$ where $j \in \phi$;
        end
        else if $\sum_\Theta \xi > \sum_\phi \xi$ then
            assign label $c$ to $i$;
        end
    end
end
```

If $n_O = n_\phi$, then we resolve the tie based on the performances of the annotators in previous assignments. We assign an *expert coder index*(ξ) to each annotator and it is updated based on the agreement of their judgments over the corpus. There are two cases when the ξ values are incremented

- If the item is assigned to a category in the gold standard data, the ξ values are incremented for those annotators who have assigned the item into that category.
- If the item is not assigned to a category in the gold standard data, the ξ values are incremented for those annotators who have not assigned the item into that category.

If n_O and n_ϕ are equal for an item, we make use of the ξ values for deciding upon the assignment of the item to the category in concern. We assign the item into that category if the combined ξ values of the annotators who have assigned the item into that category is greater than the combined ξ values of the annotators who have not assigned the item into that category, i.e.,

$$\sum_{i=1}^{n_O} \xi_i > \sum_{j=1}^{n_\phi} \xi_j$$

The algorithm for determining gold standard data is given in Algorithm 1.

5 Experimental Results

We applied the proposed A_m measure to estimate the quality of the affective corpus described in section 2. Below we present the annotation experiment followed by some relevant analysis.

5.1 Annotation Experiment

Ten human judges with the same social background participated in the study, assigning affective categories to sentences independently of one another. The annotators were provided with the annotation instructions and they were trained with some sentences not belonging to the corpus. The annotation was performed with the help of a web based annotation interface[2]. The corpus consists of 1000 sentences. Three of judges were able to complete the task within 20 days. In this paper, we report the result of applying the measure with data provided by three annotators without considering the incomplete annotations. Distribution of the sentences across the affective categories for the three judges is given in Figure 1.

5.2 Analysis of Corpus Quality

The corpus was evaluated in terms of the proposed measure. Some of the relevant observations are presented below.

- **Agreement Value:** Different agreement values related to A_m measure are given in Table 1. We present A_m values for all the annotator pairs in Table 2.

[2] http://www.mla.iitkgp.ernet.in/Annotation/index.php

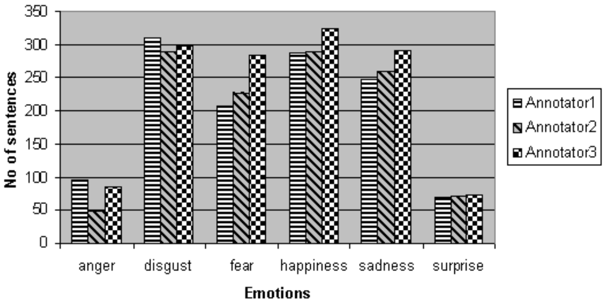

Figure 1: Distribution of sentences for three judges.

Agreement	A_m **Value**
Observed Agreement(P_o)	0.878
Chance Agreement(P_e)	0.534
A_m	**0.738**

Table 1: Agreement values for the affective text corpus.

Annotator Pair	P_o	P_e	A_m **Value**
1-2	0.858	0.526	0.702
1-3	0.868	0.54	0.713
2-3	0.884	0.531	0.752

Table 2: Annotator pairwise A_m values.

- **Agreement Study:** Table 3 provides the distribution of the sentences against individual observed agreement values. It is observed that 71.5% of the corpus belongs to [0.7 1.0] range of observed agreement and among this bulk portion of the corpus, the annotators assign 78.6% of the sentences into a single category. This is due to the existence of a dominant emotion in a sentence and in most of the cases, the sentence contains enough clues to decode it. For the non-dominant emotions in a sentence, ambiguity has been found while decoding.

Observed Agreement	**No. of Sentences**
$0.0 < A_0 \leq 0.2$	14
$0.2 < A_0 \leq 0.4$	73
$0.4 < A_0 \leq 0.7$	198
$0.7 < A_0 \leq 1.0$	715

Table 3: Distribution of the sentences over observed agreement.

- **Disagreement Study:** In Table 4, we present the category wise disagreement for all the annotator pairs. From the disagreement table it is evident that the categories with maximum number of disagreements are *anger*, *disgust* and *fear*. The emotions which are close to each other in the evaluation-activation space are inherently ambiguous. For example, anger and disgust are close to each other in the evaluation-activation space. So, ambiguity between these categories will be higher compared to other pairs. If [a b] is the pair, we count the number of cases where one annotator categorized one item into [a -] pattern and other annotator classified the same item into [- b] pattern. In Table 5, we provide the confusion between two affective categories for all annotator pairs. This confusion matrix is a symmetric one. So, we have provided only the upper triangular matrix.

  In Figure 2, we provide ambiguity counts of the affective category pairs. It can be observed that *anger*, *disgust* and *fear* are associated with three topmost ambiguous pairs.

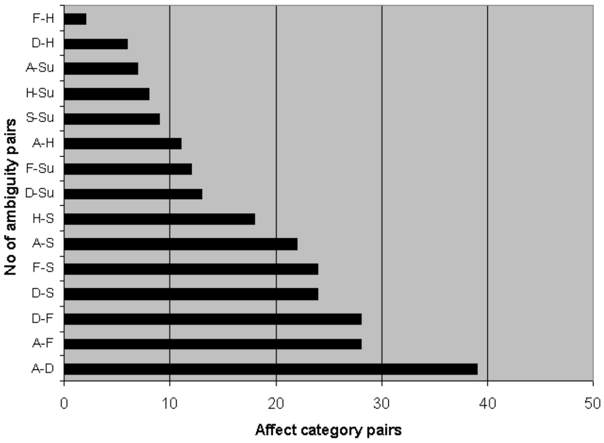

Figure 2: Category pair wise disagreement (A=Anger, D=Disgust, F=Fear, H=Happiness, S=Sadness and Su=Surprise).

5.3 Gold Standard for Affective Text Corpus

To determine the *gold standard* corpus, we have applied majority decision label based approach discussed in section 4 on the judgements provided by only three annotators. However, as the number of annotators is much less in the current study, the determined gold standard corpus may not have

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
1-2	68	94	74	64	74	45
1-3	74	86	105	57	54	45
2-3	65	49	58	22	50	20
Total	207	229	273	143	178	110

Table 4: Categorywise disagreement for the annotator pairs.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	-	39	28	11	22	7
Disgust	-	-	28	6	24	13
Fear	-	-	-	2	24	12
Happiness	-	-	-	-	18	8
Sadness	-	-	-	-	-	9
Surprise	-	-	-	-	-	-

Table 5: Confusion matrix for category pairs.

much significance. Here, we report the result of applying the gold standard determination algorithm on the data provided by three annotators. The distribution of sentences over the affective categories is depicted in Figure 3.

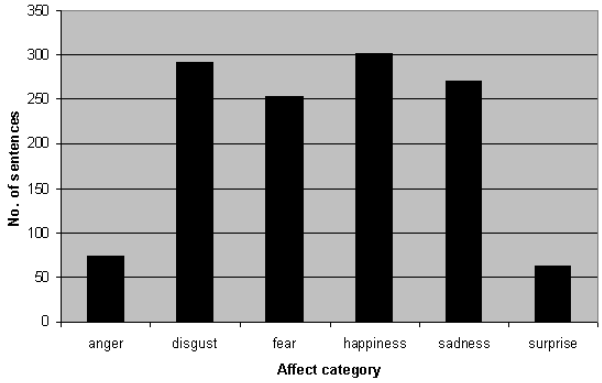

Figure 3: Distribution of sentences in gold standard corpus.

6 Conclusion and Future Work

Measuring the reliability of the affective text corpus where one single item may be classified into more than one single category is a complex task. In this paper, we have provided a new coefficient to measure reliability in multiclass annotation task by incorporating pairwise agreement in affective class pairs. The measure yields an agreement value 0.72, when applied to an annotated corpus provided by three users. This considerable agreement value indicates that the affect categories considered for annotation may be applicable to the news genre.

We are in process of collecting annotated corpus from more annotators which will ensure a statistically significant result. According to the disagreement study presented in section 5.2, confusions between specific emotions is most likely between categories which are adjacent in the activation-evaluation space. The models of annotator agreement which use weights for different types of disagreement will be interesting for future study. The direct and indirect affective sentences have not been treated separately in this study. The algorithm for determination of gold standard requires more details investigation as simple majority voting may not be sufficient for highly subjective data like emotion.

Acknowledgement

Plaban Kr. Bhowmick is partially supported by Microsoft Corporation, USA and Media Lab Asia, India. The authors are thankful to the reviewers for their detailed suggestions regarding the work.

References

Artstein, Ron and Massimo Poesio. 2008. *Inter-coder Agreement for Computational Linguistics*. Computational Linguistics.

Bruce, Rebecca F. and Janyce M. Wiebe 1999. *Rec-*

ognizing Subjectivity: A Case Study of Manual Tagging. Natural Language Engineering. 1(1):1-16.

Carletta, Jean. 1996. *Assessing Agreement on Classification Tasks: The Kappa Statistic*. Computational Linguistics. 22(21):249-254.

Carletta, Jean, Isard .A, Isard S., Jacqueline C. Kowtko, Gwyneth D. Sneddon, and Anne H. Anderson. 1997. *The Reliability of a Dialogue Structure Coding Scheme*. Computational Linguistics. 23(1):13-32.

Cohen, Jacob. 1960. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement. 20(1):37-46.

Craggs Richard and Mary M. Wood. 2004. *A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue*. Tutorial and Research Workshop, Affective Dialogue Systems. Kloster Irsee, 89-100.

Darwin, Charles. 1965. *The Expression of Emotions in Man and Animals.*. Chicago: University of Chicago Press. (Original work published 1872)

Ekman, Paul., Friesen W. V., and Ellsworth P. 1982. *What Emotion Categories or Dimensions can Observers Judge from Facial Behavior?* Emotion in the human face, Cambridge University Press. pages 39-55, New York.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. Wiley. second ed., New York.

Hagen-Zanker, Alex. 2003. *Fuzzy Set Approach to Assessing Similarity of Categorical Maps*. International Journal for Geographical Information Science. 17(3):235-249.

Hripcsak, George and Daniel F. Heitjan. 2002. *Measuring Agreement in Medical Informatics Reliability Studies*. Journal of Biomedical Informatics. 35(2):99-110.

Jung, Ho-Won. 2003. *Evaluating Interrater Agreement in SPICE-based Assessments*. Computer Standards & Interfaces. 25(5):477-499.

Krippendorff, Klaus 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications. Beverley Hills, CA.

Mieskes, Margot and Michael Strube. 2006. *Part-of-Speech Tagging of Transcribed Speech*. Proceedings of International Conference on Language Resources and Evaluation. GENOA

Ortony, Andrew and Terence J. Turner. 1990. *What's Basic About Basic Emotions?*. Psychological Review. 97(3):315-331.

Pantic, Maja and Leon Rothkrantz. 2000. *Automatic Analysis of Facial Expressions: The State of the Art*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22(12):1424-1445.

Plutchik, Robert 1980. *A General Psychoevolutionary Theory of Emotion*. Emotion: Theory, research, and experience: Vol. 1. Theories of emotion. Academic Press, New York, 3-33.

Rosenberg, Andrew, and Ed Binkowski. 2004. *Augmenting the Kappa Statistic to Determine Interannotator Reliability for Multiply Labeled Data Points*. In Proceedings of North American Chapter of the Association for Computational Linguistics. Boston, 77-80.

Scott, William A. 1955. *Reliability of Content Analysis: The Case of Nominal Scale Coding*. Public Opinion Quarterly. 19(3):321-325.

Song, Mingli, Chun Chen, Jiajun Bu, and Mingyu You. 2004. *Speech Emotion Recognition and Intensity Estimation*. Internation Conference on Computational Science and its Applications. Perugia, 406-413.

Stolcke A., Ries K., Coccaro N., Shriberg E., Bates R., Jurafsky .D, Taylor P., Martin C. Van-Ess-Dykema, and Meteer .M. 1997. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech.* Computational Linguistics. 26(3):339-371.