Predicting word sense annotation agreement

Héctor Martínez Alonso† **Anders Johannsen**† **Oier Lopez de Lacalle**‡ **Eneko Agirre**‡
†University of Copenhagen
‡University of the Basque Country
{alonso,johannsen}@hum.ku.dk {e.agirre,oier.lopezdelacalle}@ehu.eus

Abstract

High agreement is a common objective when annotating data for word senses. However, a number of factors make perfect agreement impossible, e.g. the limitations of sense inventories, the difficulty of the examples or the interpretation preferences of the annotators. Estimating potential agreement is thus a relevant task to supplement the evaluation of sense annotations. In this article we propose two methods to predict agreement on word-annotation instances. We experiment with a continuous representation and a three-way discretization of observed agreement. In spite of the difficulty of the task, we find that different levels of agreement can be identified—in particular, low-agreement examples are easier to identify.

1 Introduction

Sense-annotation tasks show less-than-perfect agreement scores. However, variation in agreement is not the result of featureless, white noise in the annotations; Krippendorff (2011) defines disagreement as *by chance*—caused by unavoidable inconsistencies in annotator behavior—and *systematic*—caused by properties of the data.

Our goal is to predict the agreement of sense-annotated examples by examining their linguistic properties. If we can identify properties predictive of low or high agreement, then we can claim that some of the agreement variation in the data is indeed systematic.

Artstein and Poesio (2008) provide an interpretation of Kripperdorff's α coefficient to describe the reliability of a whole annotation task and the way that observed agreement (A_o) is calculated for each example. Strictly speaking, the value of α only provides an indication of the replicability of an annotation task, but we propose that the difficulty of annotating a particular example will influence its local observed agreement. Thus, easy examples will have a high A_o, that will be lower for more difficult examples.

Identifying low-agreement examples by their linguistic features would help characterize contexts that make words difficult to annotate. Estimating the agreement of examples has an immediate application for data collection, as a way of estimating the proportion of examples of each difficulty level that one wants to sample. Moreover, a model of (dis)agreement can help interpret the mispredictions of a word-sense disambiguation system without requiring the data to be multiply annotated.

Observed agreement A_o is a continuous-valued variable in the unit interval and we tackle its prediction as a regression task (Section 4.1). We also experiment with a discretized version of observed agreement into low, mid and high agreement, which is predicted using classification (Section 4.2).

2 Related work

In their study, Yarowsky and Florian (2002) examine the relation between agreement variation and predictive power of word-sense disambiguation systems, which is later expanded by Lopez de Lacalle and Agirre (2015a). Our work is different in that we do not study the relation between agreement and performance, but between example properties and agreement. Martínez Alonso (2013) experiments with prediction of agreement for coarse-sense annotation.

Tomuro (2001) uses the disagreement between annotators of two English sense-annotated corpora to provide insights on the relations between synsets, and more recent studies (Jurgens, 2013; Plank et al., 2014; Jurgens, 2014; Lopez de Lacalle and Agirre, 2015b) have empirically tackled

Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 89–94,
Lisboa, Portugal, 18 September 2015.

the issue of inter-annotator disagreement as a phenomenon that is potentially informative for natural language processing. Other research efforts advocate for models of annotator behavior (Passonneau et al., 2009; Passonneau et al., 2010; Passonneau and Carpenter, 2014; Cohn and Specia, 2013).

3 Data

We conduct our study on sense-annotated datasets, keeping only the examples with at least two annotations per item. In the datasets with two annotators and one adjudicator, we disregard adjudications given their potentially different bias.

1 MASCC The English crowdsourced lexical-sample word-sense corpus from Passonneau and Carpenter (2014).

2-5 MASCE* The expert annotations for a series of English lexical-sample words from Passonneau et al. (2012), with several annotation rounds. We include the second, third and fourth round of annotation in our experiments. We use on the whole dataset (MASCEW) pooling all the rounds together, as well as on each round independently, namely MASCE2, MASCE3 and MASCE4.

6 FNTW The English Twitter FrameNet data of Søgaard et al. (2015). We treat the frame-name layer as a word-sense layer, and disregard the arguments.

7 ENSST The English supersense-annotated data of Johannsen et al. (2014).

8 EUSC The Basque lexical-sample SemCor of Agirre et al. (2006).

9 DASST The Danish supersense-annotated data of Martínez Alonso et al. (2015).

Table 1 provides the characteristics of the datasets. The annotation task can be lexical-sample (ls) or all-words (aw). The number of instances is different from the number of sentences for all-words annotation. The type of annotators can be expert (ex) or crowdsourced (cs). The α scores can differ from those reported in the datasets' documentation given our example-selection criteria. The last two columns describe the target variables of observed agreement (A_o) and the proportion of low-, mid- and high-agreement instances, cf. 3.2 for details.

3.1 Features

We define an *instance* as a sentence with a target word for annotation. If a sentence has n annotated target words, it yields n instances. For each instance, we obtain features for a word w and its syntactic parent p in a sentence s, organized in feature groups. The word identities of w and p are not included in the features to keep the models more general. Number of features are in parentheses.

Frequency(2) We calculate the frequency of w and p, scaling by $log(rank(x) + 1)^{-1}$.

Morphology (5) We consider the part-of-speech tag (POS) of w, of p, and the POS-bigram at the left and at the right of w. In order to incorporate information on inflectional complexity, we calculate which proportion of the frequency of the stem of w is covered by w, e.g. the occurrences of 'jumping' constitute 22% of the occurrences of the stem 'jump'.

Syntax (5) We calculate the number of dependents of w and p, and a bag of words for the labels of the dependents of w and p. We also include the distance from w to the root node, and the linear distance between w and p.

Context (5) We calculate the length of s in tokens, the proportion of w made up of content words, and a bag of words of the context of w, i.e. all the words of s except w. To capture context specificity, we calculate the maximum and the sum of the sentence-wise idf of each stem in s.

Sense inventory (2) We calculate the number of possible senses for w, plus an additional sense when w could be discarded from the annotation—like the tag 'O' for supersenses—or the right synset was not present in WordNet. We also calculate the sense entropy for each word following Yarowsky and Florian (2002).

We use TreeTagger (Schmid, 1994) for POS tagging and TurboParser (Martins et al., 2010) for dependency parsing, both trained on Universal Dependencies v1.1,[1] to allow cross-language feature comparison. We estimate frequencies on a 100M-word corpus for English (Ferraresi et al., 2008) and Danish (Asmussen and Halskov, 2012), and on 13M for Basque (Leturia, 2012), using Snowball stemming.

3.2 Target variable

Regression Instance-wise observed agreement (A_o) is the target variable for the regression experiments. We obtain A_o for each example by counting the pairwise matches in the annotation and dividing over the amount of pairwise combi-

[1] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/LRT-1478

Dataset	lang	inventory	task	sent	inst	ann	type	α	$A_o \pm \sigma$	L/M/H
MASCC	English	synset	ls	44.6k	44.6k	13-25	cs	.40	.14 ± .24	25/44/31
MASCEW	English	synset	ls	2.6k	2.6k	2-6	ex	.48	.07 ± .35	24/21/55
—MASCE2	English	synset	ls	1.5k	1.5k	5-6	ex	.51	.41 ± .30	21/36/43
—MASCE3	English	synset	ls	500	500	3	ex	.69	.80 ± .33	28/00/72
—MACSE4	English	synset	ls	618	618	2	ex	.63	.73 ± .44	27/00/73
ENSST	English	supersense	aw	39	326	3	ex	.67	.69 ± .36	45/00/55
FNTW	English	frame	aw	236	958	3	ex	.82	.82 ± .31	26/00/74
EUSC	Basque	synset	ls	20.6k	20.6k	2	ex	.76	.76 ± .43	24/00/76
DASST	Danish	supersense	aw	1.2k	9.5k	2	ex	.65	.67 ± .47	33/00/67

Table 1: Dataset characteristics in terms of language, sense inventory, task (al:all-words, ls:lexical sample), no. of sentences, no. of instances, no. of annotators, type of annotators (ex:expert,cs:crowdsourced), α, observed agreement and percentage of LOW/MID/HIGH agreement examples.

nations. Note that α is an aggregate measure that is obtained dataset-wise, and A_o is the only agreement measure available for individual instances.

Classification The target variable for the classification experiments is a discretization of A_o into three agreement-level classes, namely LOW, MID and HIGH. The threshold for LOW is set at $A_o \leq \frac{1}{3}$, and for HIGH at $A_o \geq \frac{2}{3}$. The MID value is only possible for datasets with more than three annotators (cf. Table 1).

4 Experiments

We use the scikit-learn[2] implementation for all learning algorithms, and train and test on 10-fold cross validation.

Regression We use L2-regularized linear regression. The baselines for regression are MEAN, where all instances receive the mean A_o of the dataset, and MEDIAN, that assigns the median A_o.

Classification We use a maximum-entropy classifier. The baselines for classification are MFC, where all instances receive most frequent class, and the two random baselines: STRA, where the assigned values are randomly selected via stratified sampling from the distribution of classes in the dataset, and UNI where values are assigned from the uniform distribution of the three labels.

4.1 Regression

Table 2 shows the results for regression in terms of mean absolute error (MAE). This metric is more suitable than root-mean-square error (RMSE) when evaluating regression in the [0,1] interval.

[2] http://scikit-learn.org/

	REGRESSION	MEAN	MEDIAN
MASCC	**0.19**	0.21	0.21
MASCEW	**0.31**	0.32	0.37
—MASCE2	0.27	0.26	0.25
—MASCE3	0.36	0.29	0.20
—MASCE4	0.46	0.40	0.27
ENSST	0.43	0.35	0.31
FNTW	0.27	0.26	0.18
EUSC	0.35	0.37	0.24
DASST	0.42	0.44	0.33

Table 2: Mean absolute error of prediction for regression and for mean and median baselines. Datasets where the system outperforms the best-performing baseline are marked in bold.

Datasets where the system outperforms both baselines are in bold.

The results for regression show that predicting instance-wise A_o is a hard task. The learnability of the task is limited by the resolution of the target variable; the only two datasets that can beat all baselines (and thus have lower MAE) have many instances, and many annotators (about 50% of the instances in MASCEW have five or more annotators). Also, size of the dataset is a relevant factor for a good estimation of A_o.

We also examine goodness of fit in terms of R^2 (determination coefficient or explained variance). R^2 does not strictly say how much agreement is systematic, but how much of the agreement variation within a dataset can be explained by the features. The only two datasets with positive R^2 are MASCC and EUSC, at .082 and 0.014 respectively. EUSC has only two annotators per instance, but it

	MAXENT	MFC	STRA	UNI
MASCC	**0.45** (0.13)	0.27	0.35	0.37
MASCEW	**0.48** (0.15)	0.39	0.39	0.35
—MASCE2	**0.39** (0.08)	0.25	0.34	0.33
—MASCE3	**0.62** (0.05)	0.60	0.57	0.53
—MASCE4	**0.63** (0.03)	0.62	0.62	0.55
ENSST	0.50 (-0.02)	0.39	0.51	0.49
FNTW	**0.71** (0.22)	0.63	0.61	0.51
EUSC	**0.68** (0.09)	0.65	0.63	0.54
DASST	**0.60** (0.11)	0.53	0.55	0.53

Table 3: Agreement prediction as classification compared against the most-frequent, stratified and uniform baseline. Datasets where the system outperforms the hardest baseline are marked in bold, error reduction in parentheses.

is a large dataset that allows mapping some properties of the features onto the variance of A_o.

The two datasets with a goodness of fit over baseline are the largest ones. This behavior indicates that the regression method suffers from the data bottleneck. Smooth estimation of continuous values might be more sensitive to data volume than estimation of discrete values, therefore we experiment with classification in the next section.

4.2 Classification

Table 3 shows the results for classification in terms of micro-averaged F_1 score. Error reduction over the hardest baseline is given in parentheses.

The ENSST dataset is the only dataset where the system cannot beat both baselines, albeit by a small margin. It is a small, all-words dataset, and the data might be too heterogenous for the model to make sense of it with only 326 instances. The F_1 scores are not very high in absolute terms, but agreement prediction is as least as hard as sense prediction.

MAE and F_1 are not comparable measures; without evaluating both on error reduction over equivalent baselines, we cannot strictly say that classification outperforms regression. Nevertheless, classification seems a promising approach.

4.3 Feature analysis

Figure 1 shows the Spearman correlation with A_o for the numeric features on two English datasets, namely MASCC and FNTW. Even though there is variation in the magnitude across datasets, we observe strong negative correlation of the sense inventory features (*z_senseentropy*, *z_nlabels*), but also for the frequency of the target word (*a_targetfreq*). Notice that these features are also colinear, and in word-sense annotation high-frequency words can be partly more difficult to annotate because they can be more polysemous.

Given these correlations, the feature repertoire we use captures better the low-agreement area of the data, but no feature has a consistently high positive correlation with agreement. That is, the predictors for low-agreement are more reliable than those for high agreement.

A possible candidate for high-agreement prediction could be the proportion of content words over the length of the context, arguably because more lexically rich context are easier to desambiguate by the annotators. This feature has a positive value for most datasets except MASCC. This property has already been noted by Passonneau et al. (2009), who mention that 'greater specificity in the contexts of use leads to higher agreement'.

Syntactic complexity is also an indicator of difficulty. Words with many dependents are often more difficult to annotate (*d_targetdeps* has a consistently negative correlation with A_o), while words with many syntactic siblings are placed in more specific contexts and are easier to annotate, giving *d_headdeps* a slight positive correlation with A_o. This behavior holds for all the English datasets except MASCC, as well as for EUSC.

We have also performed group-wise feature ablation tests on regression and classification, with similar results. Based on the contribution of single feature groups, we find that the sense-inventory group constantly outperforms the other groups, followed by the morphology group. When the sense inventory is ignored from the features, performance almost always decreases, indicating that sense inventory information is very valuable to predict agreement. However, context information is necessary to distinguish between examples of the same word (say, in a one-lemma lexical-sample dataset), where the sense-inventory features would be constant across the whole dataset.

Similarly, in the class-based experiments of Martínez Alonso (2013), certain features like plural or number of dependents are strong predictors for low agreement when annotating between the *container* and the *content* senses of words like *bowl* and *glass*. However, our datasets are ei-

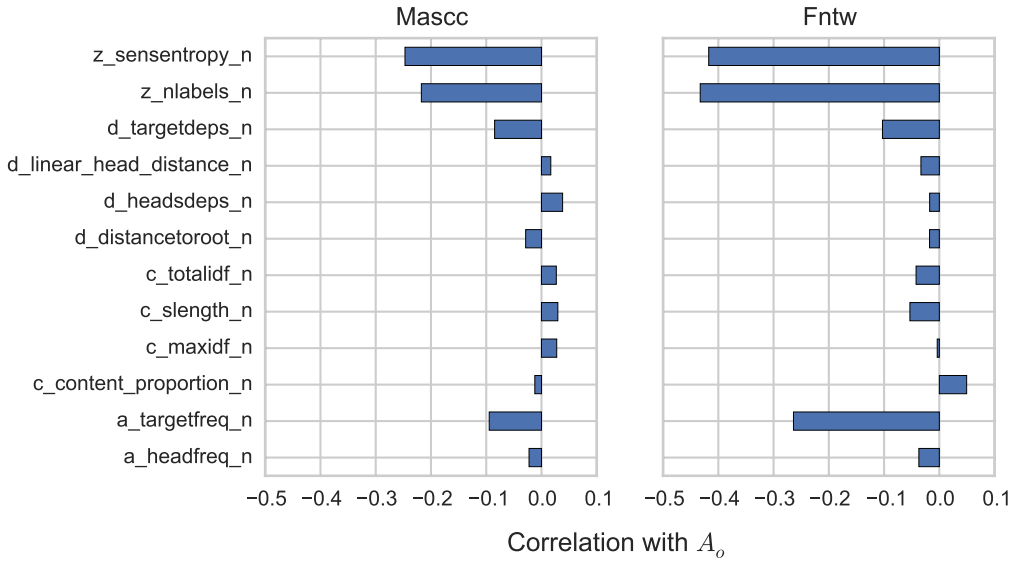

Figure 1: Correlation between the numeric-valued features and A_o for MASCC and FNTW

ther all-words or groupings of lexical-sample annotations for different words (e.g. MASC2 contains examples of *fair-j*, *know-v*, *land-n*, etc.), which means that some of the class- or lemma-dependent features might be swamped by the superposition of features from the other words.

Nevertheless, the systems do not always improve when adding context features, which suggests that there is room for improvement in capturing contextual information for sense-annotated instances.

5 Conclusions and further work

This article addresses the prediction of instance-wise agreement for sense-annotated data. We have described a method to model agreement as a continuous value, and as a set of three discrete values. We use a feature scheme that tries to give account for the lexical, morphologic and syntactic properties of the examples. We have conducted experiments on nine datasets, which comprise three languages, all-words vs. lexical-sample word annotations, and crowdsourced vs. expert annotations.

The overall conclusiveness of the study requires expanding this research to more datasets and languages, as well as further exploring the difference in annotator bias between expert and crowdsourced annotations. Our feature repertoire can be expanded with characteristics of the sense inventory in terms of sense relatedness like autohyponymy, depth in the sense ontology, or qualitative properties of the senses such abstractness. Context features can also be expanded by adding information from word sense induction and distributional models.

Moreover, if we are to examine agreement variation in full-document (as opposed to sentence-by-sentence) annotation, we suggest that document-level frequency would help concretize the meaning of a certain word, following the principle of one sense per discourse (Gale et al., 1992).

If the numeric prediction of agreement is desirable over classification, a metric like annotation entropy (Lopez de Lacalle and Agirre, 2015a) is worth considering as an alternative measure to A_o, since it an information-theoretical measure that also gives account for distribution skewness.

References

Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal, Eli Pociello, and Mikel Quintian. 2006. A methodology for the joint development of the Basque WordNet and Semcor. In *LREC*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jørg Asmussen and Jakob Halskov. 2012. The CLARIN DK Reference Corpus. In *Sprogteknologisk Workshop*.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An

application to machine translation quality estimation. In *ACL*.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. *Lexical and Computational Semantics (* SEM 2014)*.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *HLT-NAACL*.

David Jurgens. 2014. An analysis of ambiguity in word sense annotations. In *LREC*.

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112.

Igor Leturia. 2012. Evaluating different methods for automatically collecting large general corpora for Basque from the web. In *Proceedings of COLING 2012*, Mumbai, India, December. The COLING 2012 Organizing Committee.

Oier Lopez de Lacalle and Eneko Agirre. 2015a. Crowdsourced word sense annotations and difficult words and examples. *IWCS*.

Oier Lopez de Lacalle and Eneko Agirre. 2015b. A methodology for word sense disambiguation at 90% based on large-scale crowdsourcing. In *Lexical and Computational Semantics (*SEM)*.

Héctor Martínez Alonso, Anders Johannsen, Nimb Sussi, Sussi Olsen, and Bolette Sandford Pedersen. 2015. Supersense tagging for Danish. In *NODALIDA*.

Héctor Martínez Alonso. 2013. *Annotation of regular polysemy: an empirical assessment of the underspecified sense*. Ph.D. thesis, University of Copenhagen.

André FT Martins, Noah A Smith, Eric P Xing, Pedro MQ Aguiar, and Mário AT Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *EMNLP*. Association for Computational Linguistics.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *TACL*, 2:311–326.

Rebecca J Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics.

Rebecca J Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *LREC*.

Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The MASC word sense sentence corpus. In *LREC*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *ACL*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Anders Søgaard, Barbara Plank, and Héctor Martínez Alonso Alonso. 2015. Using frame semantics for knowledge extraction from Twitter. In *AAAI*.

Noriko Tomuro. 2001. Systematic polysemy and interannotator disagreement: Empirical examinations. In *First International Workshop on Generative Approaches to Lexicon*.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(04):293–310.