Supersense tagging with inter-annotator disagreement

Héctor Martínez Alonso[♠] **Anders Johannsen** **Barbara Plank**[♡]
[♡] Center for Language and Cognition, University of Groningen, The Netherlands
[♠] Univ. Paris Diderot, Sorbonne Paris Cit – Alpage, INRIA, France
hector.martinez-alonso@inria.fr,anders@johannsen.com,b.plank@rug.nl

Abstract

Linguistic annotation underlies many successful approaches in Natural Language Processing (NLP), where the annotated corpora are used for training and evaluating supervised learners. The consistency of annotation limits the performance of supervised models, and thus a lot of effort is put into obtaining high-agreement annotated datasets. Recent research has shown that annotation disagreement is not random noise, but carries a systematic signal that can be used for improving the supervised learner. However, prior work was limited in scope, focusing only on part-of-speech tagging in a single language. In this paper we broaden the experiments to a semantic task (supersense tagging) using multiple languages. In particular, we analyse how systematic disagreement is for sense annotation, and we present a preliminary study of whether patterns of disagreements transfer across languages.

1 Introduction

Consistent annotations are important if we wish to train reliable models and perform conclusive evaluation of NLP. The standard practice in annotation efforts is to define annotation guidelines that aim to minimize annotator disagreement. However, in practical annotation projects, perfect agreement is virtually unattainable. Moreover, not all of disagreement should be considered *noise* because some of it is *systematic* (Krippendorff, 2011).

The work of Plank et al. (2014a) shows that the regularity of some disagreement in part-of-speech (POS) annotation can be used to obtain more robust POS taggers. They adjust the training loss of each example according to its possible variation in agreement, providing smaller losses when a classifier training decision makes a misclassification that matches with human disagreement. For example, the loss for predicting a particle instead of an adverb is smaller than the loss for predicting a noun instead of an adverb, because the particle/adverb confusion is fairly common among annotators (Sec. 3).

In this article, we apply the method of Plank et al. (2014a) to a semantic sequence-prediction task, namely supersense tagging (SST). SST is considered a more difficult task than POS tagging, because the semantic classes are more dependent on world knowledge, and the number of supersenses is higher than the number of POS labels. We experiment with different methods to calculate the label-wise agreement (Sec. 3.1), and apply these methods to datasets in two languages, namely English and Danish (Sec. 3.2). Moreover, we also perform cross-linguistic experiments to assess how much of the annotation variation in one language can be applied to another.

2 Variation in supersense annotation

This section provides examples of reasonable disagreement in supersense annotation. We have extracted examples of disagreement from English supersense data (Johannsen et al., 2014), which we later use in our experiments. Tables 1 provides example nominal and verbal expressions, and how they have been annotated by three annotators, namely A_1–A_3.

In the first noun example, *human being* is seen by most as a two-token multiword of N.PERSON, but A_2 emphasizes the biological reading of human being when assigning senses, thus interpreting it as N.ANIMAL.

For *lightning*, we observe a disagreement across two types (N.EVENT and N.PHENOMENON) that

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 43–48,
Berlin, Germany, August 11, 2016.

	A_1	A_2	A_3
human	B-N.PERSON	B-N.ANIMAL	B-N.PERSON
being	B-N.PERSON	I-N.ANIMAL	I-N.PERSON
October	B-N.COMM.	B-N.COMM.	B-N.TIME
Iron	I-N.COMM.	I-N.COMM.	B-N.LOCATION
Range	I-N.COMM.	I-N.COMM.	I-N.LOCATION
eNews	I-N.COMM.	I-N.COMM.	B-N.COMM.
lightning	B-N.EVENT	B-N.PHEN.	B-N.PHEN.
run	V.POSS.	V.CHANGE	V.CHANGE
stop	V.MOTION	V.STATIVE	V.CHANGE
rewind	V.MOTION	V.COGNITION	V.COGNITION

Table 1: Disagreement examples. The table shows two multi-word sequences and four single words. The labels COMMUNICATION, PHENOMENON, and POSSESSION are abbreviated.

arguably have a hyponymy relation between them (phenomena being a type of event), and we consider this disagreement a consequence of the overlap in the supersense inventory. The word *thunder* shows the same disagreement.

In the case of *October Iron Range eNews*, there is disagreement on the extension of the spans of the multiword. This difference also makes A_3 provide a different semantic type to each of the three multiwords.

Even without span-size disagreements and with a slightly smaller inventory, supersense annotation for verbs is harder than for nouns. For instance, *run* is the main verb of "*He's gonna run out of money*", and even though *run* is prototypically V.MOTION, the three senses provided in Table 1 reflect the meaning of "*run out of*". In the second example, the word *stop* has full disagreement, and it even has two supersenses that seem contradictory, namely V.MOTION and V.STATIVE. This disagreement is a result of the overlap between possible annotations for *stop*.

The case of *rewind* seems more surprising, but it comes from the sentence "*Rewind the 1st time I gave you a bar of chocolate*", where *rewind* is used to mean *remember*. Both A_2 and A_3 have chosen V.COGNITION to give account for the metaphorical meaning of the verb, while A_1 has given the prototypical, literal sense of *rewind*.

3 Method

Our approach is based on the confusion-matrix cost-sensitive learning described in Plank et al. (2014a). We use a soft notion of correctness, so that the cost of making a prediction y' depends not only on whether the correct gold label y is recovered, but also on how often annotators clashed when deciding between between y and y'. The idea is to give the learner more leeway to make mistakes as long as these mistakes are the same as those made by human annotators. The learning algorithm is parameterized with a cost matrix C, where the $C_{i,j}$ is the cost of predicting j when i is the true label.

To obtain the costs, we first calculate the disagreement matrix D for each doubly-annotated dataset. An entry $D_{i,j}$ contains the probability of two annotators providing a conflicting annotation with labels i and j. High-probability entries indicate low agreement. The cost matrix is then $C_{i,j} = 1 - D_{i,j}$. In our experiments we use a structured perceptron with cost-sensitive updates as the learner.

3.1 Factorizations

While disagreement for POS is straightforward, disagreement on supersense labels can be estimated in various ways, because supersense tags contain span, POS and sense information. Supersense tags are similar to named entity tags, but using semantic types from WordNet's lexicographer files. A tag for a content word is of the form {B,I}-{POS}.{SEMANTIC-TYPE}. Function words receive the "other" tag O. Some examples of valid supersense tags are B-NOUN.PERSON, I-NOUN.PERSON or B-VERB.PERCEPTION. We abbreviate the POS block to its initial.

To account for the various kinds of information captured by the supersense tags, we use four different *factorizations*, i.e., four different ways of factoring costs into the model training. Each factorization determines when two tags are considered different in terms of applying a different loss during cost-sensitive training.

1. WHOLETAGS: disagreement over whole tags. That is, all count as disagreement if any of their parts are different, e.g., B-N.PERSON $\neq$ I-N.PERSON
2. JUSTSENSE: disagreement over the supersense, ignoring the BI prefix. That is, e.g., B-N.PERSON = I-N.PERSON, but B-N.COGNITION $\neq$ B-V.COGNITION
3. NOPOS: Only the {SEMANTIC-TYPE} block is compared, disregarding the {B,I}{POS} prefix, e.g., I-N.COGNITION = B-V.COGNITION

4. BIOPREFIX: Only the {B,I} prefix is compared, e.g., B-N.PERSON = B-V.COGNITION

3.2 Data

We use supersense data from two languages, Danish and English. For Danish, we use the SemDax corpus (Pedersen et al., 2016), a collection of supersense-annotated documents of different domains.[1] For English, we use SemCor (Miller et al., 1994) and the Twitter data presented in (Johannsen et al., 2014), RITTER-dev, RITTER-eval, and LOWLANDS. The two first Twitter data sets adds an additional layer of annotation to the corpus first introduced in Ritter et al. (2011). Table 2 provides an overview of all the individual data sets used for our supersense tagging experiments.

lang	data set	sentences	tokens
EN	SEMCOR	20132	434.7k
DA	NEWSWIRE-train	400	7k
EN	RITTER-dev	118	2.2k
EN	RITTER-eval	118	2.3k
EN	LOWLANDS	200	3k
DA	NEWSWIRE	200	3.5k
DA	BLOG	100	1.6k
DA	CHAT	200	2.9k
DA	FORUM	200	4.1k
DA	MAGAZINE	200	3.9k
DA	PARLIAMENT	200	6.2k

Table 2: Supersense tagging data sets, the first two are training data sets.

Tag inventory The English data uses the supersense inventory determined by WordNet's lexicographer files, while the Danish supersense inventory is larger, because it extends some supersenses into subtypes, e.g., N.VEHICLE, N.BUILDING and N.ARTIFACT whereas WordNet only provides N.ARTIFACT; additionally the Danish data set provides four coarse supersenses for adjectives: A.MENTAL, ADJ.PHYS, A.TIME, A.SOCIAL.

Doubly-annotated data Table 3 provides statistics on the doubly-annotated data used to calculate disagreement factorizations, including annotator agreement scores. Note that the English doubly-annotated data is considerably smaller.

[1] https://github.com/coastalcph/semdax

sample	dataset	sents	tokens	labels	A_o	κ
$\mathbb{S}_{EN}$	LOWLANDS	40	0.8k	67	0.88	0.79
$\mathbb{S}_{DA}$	NEWSWIRE	200	3.5k	71	0.68	0.53

Table 3: Statistics on the doubly-annotated data, incl. raw observed agreement A_o and Cohen's κ.

3.3 Model

Supersense tagging is typically cast as a sequential problem like POS tagging, but the class distribution is more skewed with a majority class O. We use the structured perceptron RUNGSTED, which allows cost-sensitive training.[2] We use the same feature representation as Martínez Alonso et al. (2015b), which includes information on word forms, morphology, part of speech and word embeddings. We use 5 epochs for training. All results are expressed in terms of micro-averaged F_1-score, calculated using the official CONLLEVAL.PL script from the NER shared tasks.

4 Experiments

We perform two kinds of experiments: monolingual and cross-language. For the monolingual experiments we use each of the four possible factorizations (Sec. 3.1) to train SST models with different costs on a single language. We evaluate each system against the most-frequent sense baseline (MFS), and against a regular structured perceptron without cost-sensitive training (BASELINE).

The cross-language experiments assess whether some of the disagreement information captured by the factorizations can be used cross-lingually. To study this hypothesis, we run factorized systems using $\mathbb{S}_{DA}$ (Sec. 3.1) on English, and viceversa.

Adapting $\mathbb{S}_{DA}$ to English requires projecting back to the canonical supersense inventory, namely removing the adjective supersenses and treating, e.g., all cases of NOUN.VEHICLE as N.ARTIFACT, before calculating factorizations for the different confusion matrices.

Applying the complementary process—using English disagreement information to train cost-sensitive models for Danish SST—is more involved. We have converted all the Danish data to the English SST inventory to be able to use the coarser inventory of $\mathbb{S}_{EN}$ by projecting the extended senses to their original sense. Modifying the Danish data to harmonize with $\mathbb{S}_{EN}$ has thus

[2] https://github.com/coastalcph/rungsted

lang	dataset	MFS	BASELINE	WHOLETAGS	JUSTSENSE	NOPOS	BIOPREFIX
EN	*Average*	42.51	51.36	**52.31**	**51.72**	51.13	51.13
EN	*SemCor	62.53	65.58	65.57	65.45	64.39	64.47
EN	RITTER-dev	41.54	53.44	**53.95**	52.76	52.51	52.30
EN	RITTER-eval	38.94	49.03	**49.65**	**49.97**	**49.41**	**49.42**
EN	LOWLANDS	27.11	37.38	36.93	**38.71**	**38.22**	**37.33**
DA	*Average*	33.63	40.53	39.95	**40.70**	39.94	39.08
DA	NEWSWIRE-eval	31.47	42.13	**42.21**	**42.78**	41.27	40.93
DA	BLOG	25.57	39.43	35.73	37.50	37.04	38.04
DA	CHAT	36.06	38.18	**39.12**	**38.79**	**39.81**	**38.72**
DA	FORUM	31.08	35.35	34.68	**35.45**	35.15	34.45
DA	MAGAZINE	34.28	41.97	40.91	**42.67**	**42.09**	41.44
DA	PARLIAMENT	38.57	43.04	42.81	42.84	41.32	39.20

Table 4: F_1 scores for English and Danish supersense tagging, with language-wise macro-average.

an effect on the most frequent sense baseline, because the test data is effectively relabeled.

5 Results

Table 4 shows the performance of our system compared to the MFS baseline and the non-regularized baseline that does not use factorizations. Note that our baseline structured perceptron already beats the though MFS baseline. We mark results in bold when another system beats the BASELINE. Some factorizations are more favorable for certain datasets. For instance, all factorizations improve the performance on Ritter-eval, but only WHOLETAGS aids on Ritter-dev. Over all in-language data sets, WHOLETAGS beats the macro-averaged baseline for English. However, the most reliable factorization overall is JUSTSENSE, which beats BASELINE for English and Danish.

For Danish-JUSTSENSE we observe that the adjective supersenses improve (A.MENTAL goes from 0.00 to 16.53 for a support of 15 instances, and A.SOCIAL goes from 48.87 to 56.75 for a support of 169 instances in the training data), but also other senses with much higher support improve, regardless of POS, like N.PERSON (from 49.72 to 52.66 for 951 instances) or V.COMMUNICATION (from 49.66 to 50.31 for 364 instances).

With regard to our cross-lingual investigation, only the direction of using Danish disagreement on English proves promising. Table 5 shows the results of using $\mathbb{S}_{DA}$ when training and testing on English. While JUSTSENSE still helps defeat BASELINE, using NOPOS yields better results in this setup, indicating that coarser information might be the easiest to transfer across languages. Indeed, we find that N.COMMUNICATION goes from 60.63 to 66.60 and V.COMMUNICATION goes from 71.34 to 72.05.

Unfortunately, we have not found the improvements across factorizations to be statistically significant using bootstrap test and $p < 0.05$. Some of the differences in performance for the two languages can spawn from the differences in size of the doubly-annotated sample. In fact, the amount of data in $\mathbb{S}_{DA}$ is much larger than $\mathbb{S}_{EN}$ (200 newswire sentences vs. 40 tweets).

The results indicate that there is supporting evidence that the systematicity of annotator disagreement in supersense annotation can be used for cost-sensitive training, in particular using the JUSTSENSE factorization. Notice that the improvements in Plank et al. (2014a) for tagging reach a maximum of 4 accuracy points over the regular baseline. It would be unrealistic to expect improvements of such a magnitude for SST instead of POS tagging, in particular when evaluating with label-wise micro-averaged F1 instead of accuracy.

6 Related Work

Statistical NLP has been aware of the importance of annotator bias for NLP models (Yarowsky and Florian, 2002). Ratnaparkhi and others (1996) already mentioned that annotator identity was a predictive feature for maximum-entropy POS tagging, thereby including annotator bias as a feature.

dataset	MFS	BASELINE	WHOLETAGS	JUSTSENSE	NOPOS	BIOPREFIX
Average	42.51	51.36	50.86	**51.32**	**52.52**	49.70
SemCor	62.53	65.58	64.56	64.69	**65.69**	65.06
Ritter-dev	41.54	53.44	53.04	53.52	53.42	52.31
Ritter-eval	38.94	49.03	**49.17**	**49.58**	**49.55**	**48.90**
Lowlands	27.11	37.38	38.68	37.50	37.53	32.51

Table 5: F_1s for English using cross-lingual costs calculated from $\mathbb{S}_{DA}$

Instead of training on annotator-specific data, we use disagreement to regularize over individual annotators. Tomuro (2001) has used mismatching annotations between two sense-annotated corpora to find causes of disagreement such as systematic polysemy.

Reidsma and op den Akker (2008) aim at finding ways to integrate subjective and consensual annotation in ensemble classifiers, while more recent studies (Jurgens, 2013; Aroyo and Welty, 2013; Plank et al., 2014b; Lopez de Lacalle and Agirre, 2015; Martínez Alonso et al., 2015a; Martínez Alonso et al., 2015c; Plank et al., 2015) have treated inter-annotator disagreement as potentially informative for NLP. Other research efforts advocate for models of annotator behavior (Passonneau et al., 2010; Passonneau and Carpenter, 2014; Cohn and Specia, 2013).

7 Conclusions

We presented an application of cost-sensitive learning (Plank et al., 2014a) to supersense tagging. Prior work only focused on syntactic tasks and single languages. We evaluate different factorizations of label disagreement, run monolingual experiment on languages, and attempted a cross-lingual regularization experiment.

We identify a consistent factorization (JUSTSENSE) that beats the baseline in both monolingual scenarios and in the cross-lingual scenario of using Danish annotation disagreement to train an English SST model.

We believe that capturing semantic disagreement is even more adequate for cross-lingual studies as semantics is more abstract and should better carry over to other languages. However, our investigation is only preliminary, and we would like to test the approach on further semantic tasks for which doubly-annotated data is available.

References

Lora Aroyo and Chris Welty. 2013. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *ACL*.

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

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *HLT-NAACL*, pages 556–562.

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

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

Héctor Martínez Alonso, Anders Johannsen, Oier de Lopez de Lacalle, and Eneko Agirre. 2015a. Predicting word sense annotation agreement. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, page 89.

Héctor Martínez Alonso, Anders Johannsen, Sussi Olsen, and Sanni Nimb. 2015b. Supersense tagging for danish. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, page 21.

Héctor Martínez Alonso, Barbara Plank, Arne Skjærholt, and Anders Søgaard. 2015c. Learning to parse with iaa-weighted loss. In *Proceedings of Naacl*. Association for Computational Linguistics.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, pages 240–243. 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, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *LREC*.

Bolette Sandford Pedersen, Anna Braasch, Anders Johannsen, Héctor Martínez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard, and Nicolai Sørensen. 2016. The semdax corpus–sense annotations with scalable sense inventories. In *LREC*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*.

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

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *CoNLL*.

Adwait Ratnaparkhi et al. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, USA.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Workshop on Human Judgements in Computational Linguistics, COLING*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of EMNLP*.

Noriko Tomuro. 2001. Tree-cut and a lexicon based on systematic polysemy. In *NAACL*.

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