Scoring and Classifying Implicit Positive Interpretations: A Challenge of Class Imbalance

Chantal van Son, Roser Morante, Lora Aroyo, Piek Vossen
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
{c.m.van.son, r.morantevallejo, lora.aroyo, piek.vossen}@vu.nl

Abstract

This paper reports on a reimplementation of a system on detecting implicit positive meaning from negated statements. In the original regression experiment, different positive interpretations per negation are scored according to their likelihood. We convert the scores to classes and report our results on both the regression and classification tasks. We show that a baseline taking the mean score or most frequent class is hard to beat because of class imbalance in the dataset. Our error analysis indicates that an approach that takes the information structure into account (i.e. which information is new or contrastive) may be promising, which requires looking beyond the syntactic and semantic characteristics of negated statements.

1 Introduction

Modeling inference is one of the most challenging tasks in computational linguistics, yet crucial for true natural language understanding. It requires looking beyond the literal meaning of statements and determining what exactly the author intends to convey. The position of *focus* (the new or contrastive information) in a sentence, for instance, is one pragmatic phenomenon that can affect its interpretation. The theory of alternative semantics (Rooth, 1992) assumes that the general function of focus is to evoke *alternatives*, i.e. a set of propositions potentially contrasting with the intended meaning of the sentence. A number of lexical items and constructions have been identified as focus-sensitive in English, such as *only*, *even*, *too*, counterfactual conditionals (*if*), frequency adverbs (*always*, *sometimes*, *often*) and negation. Their semantics depend on the information structure of the sentence; different choices of focus can result in different truth values. For example, Kawamura (2007) illustrates how focus contributes to the interpretation of negation in Sentence 1 below. In English, focus can be explicitly signaled prosodically, e.g. in the form of a strong pitch accent, or syntactically, e.g. by moving focused phrases to a special position in the sentence (Stevens, 2017). If no explicit signal is present, one has to rely on the context to determine which part of the sentence is in focus and, hence, what is its correct interpretation. This is a notoriously difficult task that so far has received little attention in the field.

1. (a) We do not cover PRAGMATICS in the lecture this semester.
   the lecture of this semester covers other fields of linguistics, but not pragmatics
   (b) We do not cover pragmatics IN THE LECTURE this semester.
   pragmatics is covered in the discussion sessions, homework, and exams, but not in the lecture
   (c) We do not cover pragmatics in the lecture THIS SEMESTER.
   the lecture usually covers pragmatics, but not this semester

This paper reports on a reproduction study for the purpose of understanding the performance of a system on detecting implicit positive meaning from negated statements. Such a system could support a range of Natural Language Processing (NLP) technologies that require deep understanding of language, such as Recognizing Textual Entailment (RTE) and Question Answering (QA). The research that we reproduce

Proceedings of the 27th International Conference on Computational Linguistics, pages 2253–2264
Santa Fe, New Mexico, USA, August 20-26, 2018.

is that of Blanco and Sarabi (2016), who propose an interesting methodology to automatically generate positive interpretations from negated statements and score them according to their likelihood. We reflect on the definition of the task and the contribution of the different features to the performance of their system. Our findings show that the features they propose are not able to improve upon our baseline because of class imbalance. Based on an error analysis that we perform on the output of our baseline, we discuss future lines of research that take levels of informativeness into account, which we expect to be also applicable to related problems that have to do with implicit meaning and pragmatic phenomena. The remainder of this paper is structured as follows. In Section 2, we report on related work. Section 3 describes the dataset and the task as created, defined and performed by Blanco and Sarabi (2016). In Section 4, we report on the replication of the original experiment on our test set and the ablation tests on feature combinations against the baseline. In Section 5, we redefine the task as a classification task and apply an error analysis to understand system performance against the baseline and the role of the different features in the classification task. Our error analysis leads to a discussion reported in Section 6 and plans for future work. Section 7 summarizes our conclusions.

2 Related Work

The task of scoring implicit positive meanings from negated statements was preceded by the task of detecting the *focus of negation*. This task was pioneered by Blanco and Moldovan (2011), who argued that the *scope* and *focus* of negation are crucial for a correct interpretation of negated statements. Following Huddleston and Pullum (2002), they defined scope as "the part of the meaning that is negated" and focus as "the part of the scope that is most prominently or explicitly negated." More specifically, according to Blanco and Moldovan (2011), the focus of negation gives rise to implicit positive meaning. To capture this information, they created the PB-FOC corpus, which contains annotations for the focus of negation in the 3,993 verbal negations in PropBank (Palmer et al., 2005). Candidates for focus annotation were the verb itself or any of its semantic roles and annotators were instructed to choose only one.

The PB-FOC corpus was used in the first edition of the *SEM Shared Task, which was dedicated to resolving the scope (Task 1) and focus (Task 2) of negation (Morante and Blanco, 2012). Only one team (Rosenberg and Bergler, 2012) participated in the Focus Detection task. Their system, called CLaCs NegFocus, is rule-based and consists of three components: the identification of explicit negation cues (*not*, *nor* and *never*), the detection of the syntactic scope of negation, and the detection of the focus of negation using a set of four syntactic heuristics (e.g. "if an adverb directly precedes the negated verb and is connected through an *advmod* dependency relation to the negated verb, this adverb is annotated as the focus"). They report an F-measure of 0.584. Blanco and Moldovan (2013) report an F-measure of 0.641 with their system FOC-DET, which is trained with bagging over standard C4.5 decision trees using a set of features derived from gold syntactic annotation and semantic role labels. Whereas both CLaCs NegFocus and FOC-DET use only information from within the sentence, Zou et al. (2015) argue that contextual discourse information plays a critical role in negation focus identification and propose a word-topic graph model that uses this contextual information from both lexical and topical perspectives. They report an accuracy of 69.39 on PB-FOC.

PB-FOC has given rise to various alternative approaches to the annotation of focus and different definitions of the corresponding task. Blanco and Moldovan (2012) introduce the concept of granularity of focus and add fine-grained foci on top of the course-grained foci in PB-FOC; whereas coarse-grained focus includes all words belonging to a semantic role, fine-grained focus comprises fewer words within the semantic role (e.g. *We didn't get [an offer for more than $40]*$_{\text{FOCUS}}$), allowing for more specific implicit positive interpretations ("we got something, but not an offer for more than $40" versus "we got an offer for $40 or less"). Anand and Martell (2012) evaluated the annotations of PB-FOC, arguing that positive interpretations resulting from scalar implicatures and neg-raising predicates should be separated from those (indirectly) resulting from focus, and reannotated the corpus by using an alternative annotation approach that relies on relevant questions under discussion (QUDs).

Another criticism on PB-FOC was raised by Blanco and Sarabi (2016), who point out that the guidelines required the annotators to choose one semantic role as the focus, prioritizing "the one that yields the most meaningful implicit [positive] information" in case of multiple candidates, but that it is not specified what "most meaningful" means. Therefore, they designed a new annotation task where several positive interpretations per negation (automatically generated by manipulating semantic roles) are scored according to their likelihood (see Section 3 for more details). Similar to the distinction between fine-grained and coarse-grained foci, Sarabi and Blanco (2016) propose to extract and score more fine-grained positive interpretations by manipulating syntactic dependencies instead of semantic roles. Sanders and Blanco (2016) further extend this approach of generating and scoring implicit positive interpretations by applying it to modal constructions.

3 Dataset and Task

The dataset created by Blanco and Sarabi (2016) contains 1,888 positive interpretations generated from 600 negated verbs in OntoNotes 5.0 (Hovy et al., 2006). These positive interpretations are automatically generated by first converting the negated statement into its positive counterpart by (1) removing the negation mark, (2) removing auxiliaries, expanding contractions and fixing third-person singular and past tense, and (3) rewriting negatively-oriented polarity-sensitive items (e.g. *anyone* becomes *someone*). From this positive counterpart, positive interpretations are generated by rewriting each semantic role or the (originally negated) verb. For example, ARG0-ARG4 are rewritten as *someone / some people / something*, and ARGM-TMP is rewritten as *at some point of time*. To illustrate, this results in the three positive interpretations listed for Sentence 2.

2. [The proper climatic conditions]$_{\text{ARG1}}$ don't [exist]$_{\text{V}}$ [in many places in the world]$_{\text{ARGM-LOC}}$.
   (a) The proper climatic conditions {some verb / action / state} in many places in the world.
   but not 'exist'
   (b) {someone / some people / something} exist in many places in the world.
   but not 'The proper climatic conditions'
   (c) The proper climatic conditions exist {somewhere}.
   but not 'in many places in the world'

On average, 3.15 positive interpretations are generated per negation, each of which is manually scored with a value between 0 and 5. A positive interpretation with a score of 5 is deemed very plausible given the context (i.e. the previous and next sentence), whereas a score of 0 indicates that this interpretation is highly implausible. Inter-annotator agreement is calculated using Pearson's correlation and is reported to be 0.761. The task is defined as a regression task, where each positive interpretation becomes an instance for which the system should produce a score.

3.1 Data releases

The dataset of scored positive interpretations is not publicly available, but the authors provided us with all their annotations. Blanco and Sarabi (2016) have used the CoNLL-2011 Shared Task distribution of OntoNotes[1] (Pradhan et al., 2011). In order to use the annotations for learning, both OntoNotes 5.0 and the CoNLL-2011 Shared Task release are required, since the latter contains all annotations but not the underlying words. The organizers of this shared task provide scripts and instructions to merge the annotations with the words in OntoNotes. Originally, the data was meant to be merged with the data from OntoNotes 4.0,[2] but it can also be used with OntoNotes 5.0,[3] which is the final release of the OntoNotes project. The resulting `.conll` files contain a merged representation of all the OntoNotes layers in CoNLL-style tabular format with one line per token, and with multiple columns specifying the annotations for each token. This format is similar to the CoNLL-formatted release of OntoNotes 5.0.[4] However, there are two important differences between the two CoNLL-formatted releases. First, for

[1] http://conll.cemantix.org/2011

[2] https://catalog.ldc.upenn.edu/ldc2011t03

[3] https://catalog.ldc.upenn.edu/ldc2013t19

[4] https://github.com/propbank/propbank-release

Source	Feature description
verb	Word form and part-of-speech tag of verb
sem_role	Semantic role label of sem_role (sem_role_label)
	Number of tokens in sem_role
	Word form and part-of-speech tag of head of sem_role
	Syntactic node of sem_role, its parent and left and right siblings in the parse tree
verb-sem_role	Whether verb occurs before or after the sem_role in the negated statement
	Syntactic node of lowest common ancestor of verb and sem_role
	Syntactic paths from verb to sem_role
verbarg-struct	Flags indicating whether verb has each possible semantic role
	Semantic role labels of the first and last roles of verb
	Syntactic nodes and heads of each semantic role attaching to verb

Table 1: Features used by Blanco and Sarabi (2016, p. 1438)

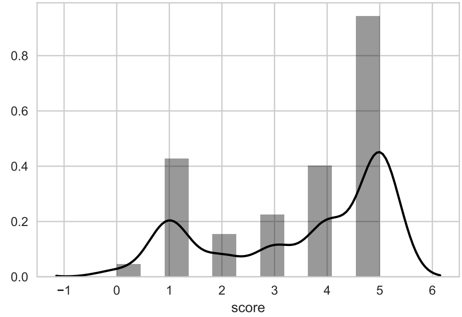


Figure 1: Overall distribution of scores in the complete dataset.

	count	mean	std	min	max
role_label					
ARG0	316	4.02	1.09	0.0	5.0
ARG1	416	4.53	1.01	0.0	5.0
ARG2	82	4.37	1.35	0.0	5.0
ARG3	1	5.00	NaN	5.0	5.0
ARG4	2	5.00	0.00	5.0	5.0
ARGM-ADV	80	2.40	1.65	0.0	5.0
ARGM-CAU	12	2.50	2.07	0.0	5.0
ARGM-DIR	7	2.43	2.44	0.0	5.0
ARGM-EXT	9	3.56	1.74	1.0	5.0
ARGM-LOC	16	3.62	1.75	0.0	5.0
ARGM-MNR	28	4.54	1.14	0.0	5.0
ARGM-PNC	3	5.00	0.00	5.0	5.0
ARGM-PRP	1	2.00	NaN	2.0	2.0
ARGM-TMP	65	4.03	1.48	0.0	5.0
V	472	2.20	1.40	0.0	5.0

Table 2: Statistics of scores per semantic role in the train set. The mean scores per role are used as a baseline.

CoNLL-2011 some of the documents were split into smaller parts to reduce the complexity of coreference annotation, which increases with the length of a document. Second, disfluencies in conversation data (e.g. restarts, hesitations), which are marked with a special EDITED phrase tag in the original OntoNotes parses, were removed. The splitting of documents and removal of disfluencies result in different sentence and token identifiers, since sentence numbering restarts for each document part and token count excludes disfluent tokens. Therefore, the CoNLL-2011 release is required for resolving the identifiers used in the positive interpretations dataset. However, the CoNLL-2011 release does not provide gold annotations for the complete dataset; for the test set used for this shared task, only automatic annotations are provided. The gold annotations required to replicate the experiment by Blanco and Sarabi (2016) were provided to us by the authors instead.

4 Regression Task: Scoring Positive Interpretations

The system for scoring the positive interpretations as reported in (Blanco and Sarabi, 2016) is not publicly available. However, since the authors provided us with the annotations and the required OntoNotes data, we were able to replicate their experiment for further analysis.

4.1 Experimental set-up

Simulating the approach reported by Blanco and Sarabi (2016), we trained a Support Vector Machine (SVM) for regression with RBF kernel using scikit-learn (Pedregosa et al., 2011) with the set of features

	Blanco & Sarabi	ANONYMIZED		
	Pearson's r	Pearson's r	Spearman's ρ	RMSE
BASELINE (MEAN)		**0.679**	**0.625**	**1.239**
{semrole_label}	**0.603**	**0.656**	0.639	**1.282**
{semrole}		0.632	**0.640**	1.383
{verb-semrole}		0.603	0.595	1.357
{verbarg-struct}		0.148*	0.140*	1.747
{verb}	-0.025	0.078*	0.061*	1.760
{semrole_label, verbarg-struct}		**0.655**	0.631	1.562
{semrole_label, verb-semrole}		0.640	0.619	**1.299**
{semrole, verb-semrole}		0.639	**0.653**	1.367
{verb, semrole_label}		0.633	0.637	1.383
{verb-semrole, verbarg-struct}		0.632	0.615	1.540
{verb, semrole}	**0.630**	0.622	0.634	1.391
{semrole, verbarg-struct}		0.605	0.623	1.431
{verb, verb-semrole}		0.585	0.605	1.436
{verb, verbarg-struct}		0.141*	0.136*	1.749
{verb, semrole_label, verbarg-struct}		**0.655**	0.628	1.591
{semrole_label, verb-semrole, verbarg-struct}		0.642	**0.640**	1.436
{verb, semrole, verb-semrole}	**0.627**	0.631	0.643	**1.387**
{semrole, verb-semrole, verbarg-struct}		0.625	0.633	1.404
{verb, semrole_label, verb-semrole}		0.618	0.615	1.397
{verb, semrole, verbarg-struct}		0.597	0.619	1.440
{verb, semrole_label, verb-semrole, verbarg-struct}		**0.646**	**0.642**	1.452
{verb, semrole, verb-semrole, verbarg-struct}	**0.642**	0.620	0.631	**1.417**

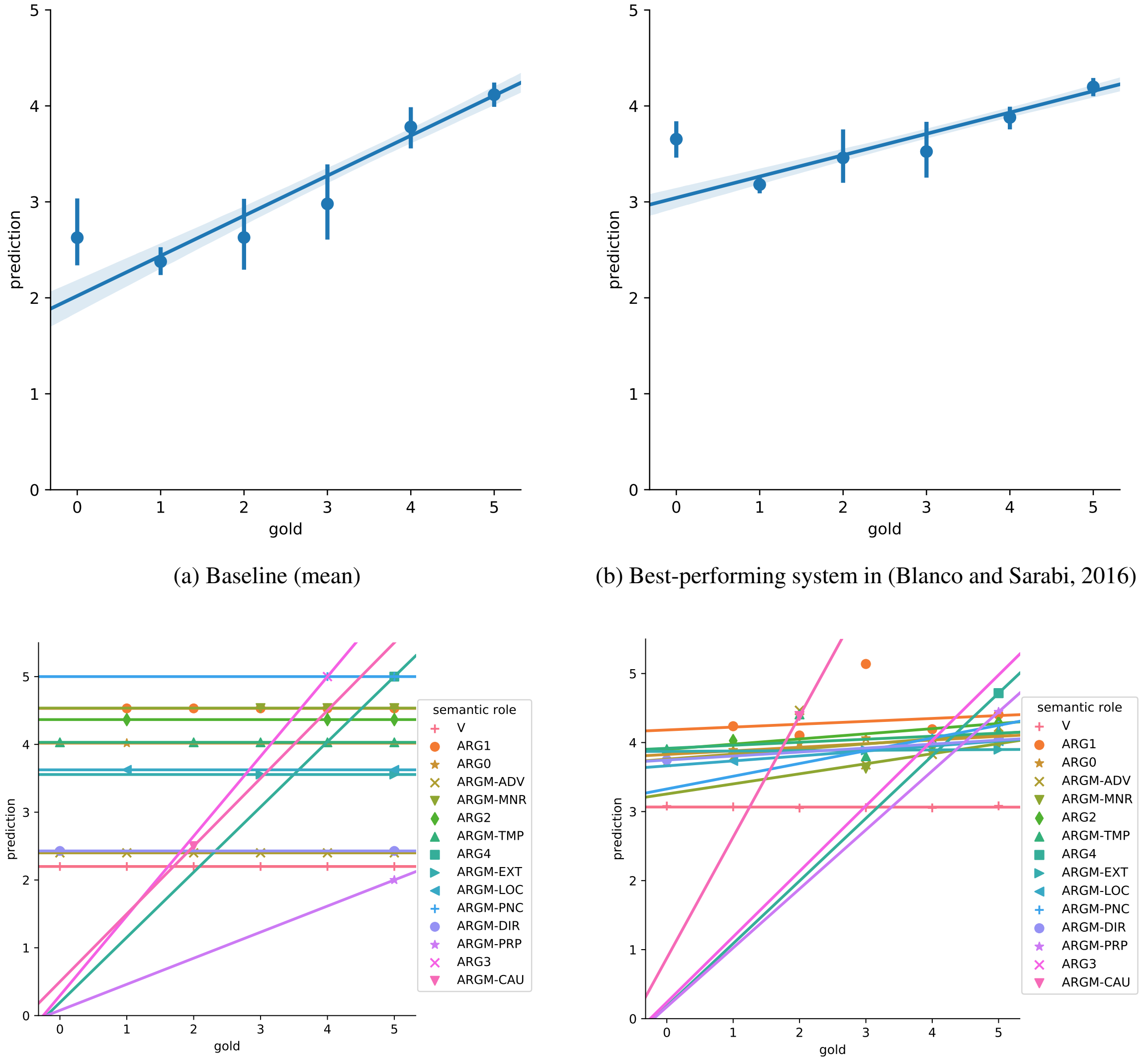
Table 3: Results of the Support Vector Machine (SVM) for regression with RBF kernel with different feature sets compared to those reported in (Blanco and Sarabi, 2016). All were statistically significant ($p < 0.001$), except those indicated with an asterisk.

presented in Table 1. Whereas Blanco and Sarabi (2016) report on results obtained with features extracted from gold-standard and predicted linguistic annotations, we only report on those obtained from the gold-standard annotations in the CoNLL-2011 dataset. We compare the results to a baseline system that simply takes the mean score of the semantic role from which the positive interpretation was generated as calculated over our train set (Table 2). Blanco and Sarabi (2016) evaluate their results using Pearson's correlation (r). However, we question the appropriateness of this evaluation measure, since Pearson's r assumes that the data is continuous and (at least approximately) normally distributed. Whereas the predicted scores are continuous indeed, the gold scores are rather ordinal of nature. Moreover, the data is not normally distributed, as can be seen in Figure 1. Therefore, we report our results using two additional evaluation measures: Spearman's correlation (ρ) with correction for ties, which can be used when the assumptions of Pearson's r are not met, and Root Mean Square Error (RMSE), which is the standard deviation of the residuals (prediction errors). RMSE is a measure of how spread out the residuals are; this way, it informs how concentrated the predicted scores are around the line of best fit. Lower values of RMSE indicate better fit, and since it is an absolute measure with the same units as the score being estimated, the values should be interpreted in the same range, in our case, between 0 and 5.

Unfortunately, due to time constraints, the authors were not (yet) able to provide us with the train/test splits as used in their paper. Therefore, we created our own test set using the same approach: we used a 80/20 split and made sure that all positive interpretations belonging to one negated verb were assigned to either the train or the test set.

4.2 Results

Table 3 presents the results of our experiments with the different combinations of feature sets on the regression task. We observe that, depending on the evaluation measure, different systems perform best. We also observe that our results are mostly comparable to those reported in (Blanco and Sarabi, 2016),



(a) Baseline (mean)

(b) Best-performing system in (Blanco and Sarabi, 2016)

(c) Baseline (mean)

(d) Best-performing system in (Blanco and Sarabi, 2016)

Figure 2: The upper figures show regression lines with a 95% confidence interval. The observations are collapsed in bins to plot an estimate of central tendency. The lower figures show regression lines per semantic role.

although just using the semrole_label as feature shows better results and is even the second-best system after our baseline, if we consider Pearson's r and the RMSE scores. This contradicts the finding of Blanco and Sarabi (2016), who found that considering all features yields the best performance. However, we emphasize that we tested on a different test set than Blanco and Sarabi (2016), so it is possible that this accounts for the differences in results. If we look at Spearman's ρ, the feature combination {semrole, verb-semrole} performs best, followed by {verb, semrole_label, verb-semrole, verbarg-struct}.

Figures 2a-2d visualize the regression lines for the baseline system and for the system that was reported as best-performing in (Blanco and Sarabi, 2016). We show both the overall tendency as the relation between the gold and the predicted scores per semantic role. As expected, Figure 2c shows flat lines for each semantic role, corresponding to the mean scores of each, except for those semantic roles that occur only once in the test set (ARG3, ARGM-CAU and ARGM-PRP) and ARG4, which is correctly scored as 5 in all cases. In contrast, Figure 2d shows more variation within most semantic roles (indicated by their slope), but in general, the predicted scores are too high, as can also be concluded from the overall overview in Figure 2b. In fact, the lowest predicted score according to this system is 2.96. The baseline seems to account a bit better for the lower and middle scores (the lowest predicted score is 2).

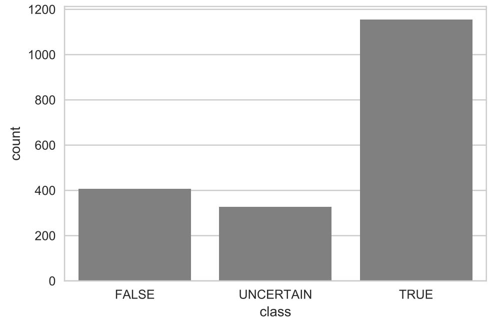


Figure 3: Overall distribution of the three classes in the complete dataset.

role_label	FALSE	UNCERTAIN	TRUE	total
ARG0	19	45	**252**	316
ARG1	14	26	**376**	416
ARG2	7	4	**71**	82
ARG3	-	-	**1**	1
ARG4	-	-	**2**	2
ARGM-ADV	**33**	24	23	80
ARGM-CAU	**6**	2	4	12
ARGM-DIR	**4**	-	3	7
ARGM-EXT	2	1	**6**	9
ARGM-LOC	3	3	**10**	16
ARGM-MNR	1	2	**25**	28
ARGM-PNC	-	-	**3**	3
ARGM-PRP	-	1	-	1
ARGM-TMP	8	9	**48**	65
V	**224**	163	85	472

Table 4: Statistics of the three classes per semantic role in the train set. The most frequent classes per role (in bold) are used as a baseline.

5 Classification Task: Classifying Positive Interpretations as TRUE or FALSE

In a second experiment, we redefine the task as a classification task. We argue that the scores assigned to each positive interpretation in the dataset reflect the confidence of the annotators rather than what we are interested in predicting. In other words, in the end a binary classification system that can predict whether some positive meaning is implied in a negated statement or not would suffice in most real-world applications. Instead of predicting a score between 0 and 5, one option is thus to redefine the task as a binary classification problem where each positive interpretation is to be classified as either TRUE or FALSE. The original annotated scores can easily be divided: all scores between 0-2 become FALSE, and all scores between 3-5 become TRUE. However, intuitively, it is clear that the scores 2 and 3 are somewhat problematic in the sense that they present the middle cases that less clearly belong to either class. Therefore, these two scores could potentially make up their own class of UNCERTAIN. As can be seen in Figure 3, the UNCERTAIN class comprises a considerably large portion of the complete dataset.

5.1 Experimental set-up

We use the same sets of features to train an SVM with RBF kernel, but now for classification. We apply it to both the binary classification task (TRUE/FALSE) and the tertiary classification task (TRUE/FALSE/UNCERTAIN). The results are evaluated using precision, recall and F1-measure. We use weighted averaging for all scores, which accounts for class imbalance by computing the average of binary metrics in which each class' score is weighted by its presence in the true data sample. Our baseline system takes the most frequent class of the semantic role from which the positive interpretation was generated (corresponding to the classes in bold in Table 4 in both tasks).

5.2 Results

We summarize our results on both the tertiary and the binary classification tasks in Table 5. None of the feature combinations were able to beat the most-frequent baseline, which achieves an F1-measure of 0.725 with three classes and 0.840 with two classes. The differences with the second-best system, which uses only semrole_label as a feature, are small (0.721 and 0.834 respectively), but adding more features only appears to hurt performance. This is mostly in line with what we observed in the regression task. We analyzed the precision, recall and F1-measure per class for each of the feature combinations and observe that for most of them, the FALSE class performs 0.0 on each in both tasks, except for the better performing ones (indicated with bold scores in Table 5). Moreover, the UNCERTAIN class always performs 0.0 in the tertiary task, including the baseline. In the binary task, the baseline achieves 0.65 / 0.89 / 0.75 on precision, recall and F1-measure on the FALSE class and 0.95 / 0.81 / 0.87 on the TRUE class. In the tertiary task, the respective scores are 0.54 / 0.91 / 0.68 (FALSE) and 0.91 / 0.85 / 0.88 (TRUE).

	Tertiary Classification			Binary Classification		
Features	Precision	Recall	F1	Precision	Recall	F1
BASELINE (MOST FREQUENT)	**0.711**	**0.762**	**0.725**	**0.864**	**0.833**	**0.840**
{semrole_label}	**0.702**	**0.759**	**0.721**	**0.853**	**0.828**	**0.834**
{verb-semrole}	0.681	0.759	0.716	0.832	0.825	0.828
{semrole}	0.619	0.714	0.663	0.514	0.717	0.599
{verb}	0.424	0.651	0.513	0.514	0.717	0.599
{verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{semrole_label, verb-semrole}	**0.681**	**0.759**	**0.716**	**0.832**	**0.825**	**0.828**
{verb, verb-semrole}	**0.681**	**0.759**	**0.716**	**0.832**	**0.825**	**0.828**
{semrole, verb-semrole}	**0.681**	**0.759**	**0.716**	0.821	0.817	0.819
{verb, semrole_label}	**0.681**	**0.759**	**0.716**	0.514	0.717	0.599
{verb, semrole}	0.424	0.651	0.513	0.514	0.717	0.599
{verb, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{semrole, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{semrole_label, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{verb-semrole, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{verb, semrole_label, verb-semrole}	**0.681**	**0.759**	**0.716**	**0.832**	**0.825**	**0.828**
{verb, semrole, verb-semrole}	0.653	0.735	0.691	0.514	0.717	0.599
{verb, semrole, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{verb, semrole_label, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{verb, verb-semrole, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{semrole, verb-semrole, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{semrole_label, verb-semrole, verbarg-struct}	0.424	0.651	0.513	0.514	0.717	0.599
{verb, semrole, verb-semrole, verbarg-struct}	**0.424**	**0.651**	**0.513**	**0.514**	**0.717**	**0.599**
{verb, semrole_label, verb-semrole, verbarg-struct}	**0.424**	**0.651**	**0.513**	**0.514**	**0.717**	**0.599**

Table 5: Results of the Support Vector Machine (SVM) for classification with RBF kernel with different feature sets on the classification task with three classes and with two classes.

5.3 Error analysis

The results on the classification task can easily be explained if we have another look at Figure 3 and Table 4. ARG0, ARG1 and V are by far the most frequent cases in the dataset and their classes are very imbalanced. ARGM-ADV, ARGM-CAU and ARGM-DIR are the only roles for which FALSE is the most frequent class and the classes are somewhat balanced, but these roles are not as well represented in the dataset. As expected, our error analysis on the results of the baseline show that the positive interpretations generated from the verb itself comprise the largest portion of classification errors: 39 out of 63 (62%) were incorrectly predicted as FALSE. This is followed by ARGM-ADV (16%, incorrectly predicted as FALSE), ARG1 (8%, TRUE), ARGM-TMP (5%, TRUE) and ARG0 (3%, TRUE). ARG2, ARGM-DIR, ARGM-LOC and ARGM-PRP were only misclassified once. From these results, we learn that the challenge of this task is to make a system perform well for the minority classes of each role. In other words, we need to know when the verb or ARGM-ADV results to TRUE and when the other roles result to FALSE.

We discuss some examples of misclassification. According to the gold annotations, the positive interpretation generated from the verb *create* in Sentence 3 results to TRUE. We argue that this is because the sentence contains *contrastive focus*. That is, *create* contrasts with another verb mentioned in the previous context: *enrich*. Evidence is provided by the subject being shared between the verbs, i.e. *a diversity of cultures*. Generally speaking, implicit positive meanings involving contrasting actions seem to be more easily evoked when the verb has a strong antonym, even if this is not explicitly mentioned in the context. For example, verbs like *win* (*lose*) or *like* (*dislike*) are more likely to give rise to alternatives than verbs that do not have a clear opposite meaning, such as *write*.

3. [A diversity of cultures]$_{\text{ARG0}}$ will only enrich a person, not **create** [a crisis of identity]$_{\text{ARG1}}$.
   A diversity of cultures will only enrich a person, {some verb / action / state} a crisis of identity, but not 'create'

Another observation in the TRUE positive interpretations generated from the verb, is that many of them contain coreferential expressions. For example, all semantic roles belonging to the negated verb in Sentence 4 consist of pronouns, indicating that the agent and patient have already been introduced in the previous context. The acceptability of this positive interpretation seems to be a result of *new information focus*, since the only 'new' information is the action expressed by the verb. This makes it more likely that the negation of the action is what the author intended to convey.

4. But [he]$_{\text{ARG0}}$ didn't **win** [it]$_{\text{ARG1}}$.
But he {some verb / action / state} it, but not 'won'

Sentence 5, on the other hand, illustrates how information that is already introduced in the previous context is less likely to be in the pragmatic scope of negation. The ARG1, *the advance*, is coreferential with *gained* in the preceding context; the new, relevant information is the lack of participation of *many issues* in this advance. Therefore, in this case, the positive interpretation generated from this semantic role is FALSE instead of the more frequent TRUE class associated with ARG1.

5. But while the Composite gained 1.19, to 462.89, [many issues]$_{\text{ARG0}}$ didn't **participate** [in the advance]$_{\text{ARG1}}$.
But while the Composite gained 1.19, to 462.89, many issues participated {in someone / some people / something}, but not 'in the advance'

6 Discussion and future work

The relation between information structure and inference is an interesting research area that needs further exploration in computational linguistics. Blanco and Moldovan (2011) laid some important groundwork for deriving implicit positive meaning from negations by trying to detect the (single) focus of negation. However, as correctly noted by Anand and Martell (2012), implicit positive meaning only indirectly results from focus. This is partially because, as Kawamura (2007) explains, negation does not *require* a focused element (in contrast to focus-sensitive elements such as *only*, *even* and *either*) and at the same time induces ambiguity in focus constructions. That is, negation without focus can still be grammatical; its absence simply produces a different interpretation. If the sentence *does* contain focus, negation yields a focus-associated reading (6a) and a non-associated reading (6b), as Kawamura (2007) illustrates with the example below. This ambiguity has been regarded as a result of the scope interaction of focus and negation (Jackendoff, 1972; Krifka, 2006).

6. John didn't criticize Mary [at the conference]$_{\text{FOCUS}}$.
   (a) John criticized Mary. It was not at the conference.
   (b) John did something other than criticizing Mary. It happened at the conference.

Therefore, we think Blanco and Sarabi (2016) improved upon the task by scoring each and every potential positive interpretation generated from the negation. The question remains, however, whether we really need to score positive interpretations on a scale from 0 to 5. We believe that, from an application perspective, classifying them as TRUE/FALSE (and optionally UNCERTAIN) is sufficient.

We have shown that class imbalance in the dataset is the reason why our baseline performs so well on the task, and that –even though there is no one-to-one correspondence between focus and implicit positive meaning– knowing which part of the proposition is new or contrastive seems to play a crucial role in correctly predicting the infrequent classes. We therefore argue that an approach to detecting implicit meaning should aim to include this pragmatic layer of information. The features described in this paper only partially capture this layer since they rely on explicit signals within the sentence (i.e. by determining which information is present and in what position), but they do not take context into account. However, we need more annotated data and more experiments to confirm our hypothesis.

Special cases of negation that we would like to highlight here are those that contain negatively-oriented polarity-sensitive items, such as *anyone*, *anything* or *either*. In the current definition of the task, these items are rewritten to positively-oriented polarity-sensitive items (e.g. *anyone* becomes *someone*, *anything* becomes *something*) before generating the positive interpretations. For example, Sentence 7 would result in the interpretations 7a and 7b with this procedure. While 7a may be classified as UNCERTAIN (or, less likely, as TRUE), 7b will always be FALSE. Moreover, neither interpretations cover the actual meaning of the sentence. Instead, it should be interpreted as *"Kim read nothing"*, which in turn brings us to the next question: how to deal with other markers of negation, such as *nothing* or *nobody*? We suggest treating sentences with these kind of markers of negation or with negatively-oriented polarity-sensitive items both as special cases that probably do not need any annotation, since their correct interpretation directly follows from the meaning of the words.

7. Kim didn't read anything.
   (a) {*someone / some people / something*} *read something (but not Kim)*
   (b) *Kim read* {*something*}

Finally, while the task explored in this paper restricts itself to negation, it could possibly be extended to include also inferences resulting from focus in sentences without negation or with other focus-sensitive elements. As Kawamura (2007), among others, describes, the semantic effect of focus depends on the type of focus-sensitive elements present in the sentence. As noted above, some focus-sensitive elements require a focused element for their interpretation, while others do not. In addition, in sentences without focus-sensitive elements, focus-shift does not affect the truth conditions (even though intuitively the interpretations are different), whereas the truth conditions of sentences with a focus-sensitive element such as *only* do change. That is, if John reads something other than books in the bathroom, Sentence 8a and 8b can both be TRUE, while the focus-shift causes Sentence 9a to be FALSE.

8. (a) John reads [books]$_{\text{FOCUS}}$ in the bathroom.
   (b) John reads books [in the bathroom]$_{\text{FOCUS}}$.

9. (a) John only reads [books]$_{\text{FOCUS}}$ in the bathroom.
   (b) John only reads books [in the bathroom]$_{\text{FOCUS}}$.

This leaves us with many opportunities for future work to further investigate the inferences and truth conditions of sentences in relation to the interaction between focus and focus-sensitive elements.

7 Conclusion

We reimplemented a system for scoring implicit positive interpretations from negated statements for the purpose of understanding its performance, the task itself and the role that different features play. We have argued that the original task could be redefined as a classification task, where each positive interpretation is to be classified as TRUE or FALSE (optionally, with a third class of UNCERTAIN representing the middle cases). We showed that a simple baseline that takes the mean over the scores or the most frequent class per semantic role is hard to beat in both the regression and the classification tasks. Our error analysis indicated that improvements can only be achieved with a system that is able to account for the infrequent classes, and we hypothesize that modeling informativeness (e.g. new information or contrastive focus) may be useful in this respect. This would require looking beyond the syntactic and semantic characteristics of the proposition and taking the broader context into account. Our code is publicly available at `https://github.com/cltl/positive-interpretations`.

Acknowledgements

This research is supported by the Amsterdam Data Alliance in the QuPiD project and the Netherlands Organization for Scientific Research (NWO) via the Spinoza-prize awarded to Piek Vossen in the project "Understanding Language by Machines" (SPI 30-673, 2014-2019). We would like to thank Zahra Sarabi and Eduardo Blanco for sharing their data and for their willingness to answer our many questions. We are also grateful to the anonymous reviewers for their valuable comments.

References

Pranav Anand and Craig Martell. 2012. Annotating the focus of negation in terms of Questions Under Discussion. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 65–69.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 581–589.

Eduardo Blanco and Dan Moldovan. 2012. Fine-grained focus for pinpointing positive implicit meaning from negated statements. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 456–465, Montréal, Canada, June. Association for Computational Linguistics.

Eduardo Blanco and Dan Moldovan. 2013. Retrieving implicit positive meaning from negated statements. *Natural Language Engineering*, 20(4):501–535.

Eduardo Blanco and Zahra Sarabi. 2016. Automatic generation and scoring of positive interpretations from negated statements. In *Proceedings of NAACL-HLT*, pages 1431–1441.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Rodney Huddleston and Geoffrey K Pullum. 2002. *The Cambridge Grammar of English Language*. Cambridge: Cambridge University Press.

Ray S Jackendoff. 1972. *Semantic interpretation in generative grammar*. M.I.T. Press,, Cambridge, Mass.

Tomoko Kawamura. 2007. *Some interactions of focus and focus sensitive elements*. State University of New York at Stony Brook.

Manfred Krifka. 2006. Association with Focus. In Valerie Molnar and Susanne Winkler, editors, *The Architecture of Focus*, pages 105–136. Mouton de Gruyter.

Roser Morante and Eduardo Blanco. 2012. * SEM 2012 Shared Task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 265–274, Montréal, Canada.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon.

Mats Rooth. 1992. A theory of focus interpretation. *Natural language semantics*, 1(1):75–116.

Sabine Rosenberg and Sabine Bergler. 2012. UConcordia: CLaC negation focus detection at *SEM 2012. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 294–300.

Jordan Sanders and Eduardo Blanco. 2016. Automatic extraction of implicit interpretations from modal constructions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA.

Zahra Sarabi and Eduardo Blanco. 2016. Understanding negation in positive terms using syntactic dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA.

John Scott Stevens. 2017. The pragmatics of focus. In *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.

Bowei Zou, Qiaoming Zhu, and Zhou Guodong. 2015. Unsupervised negation focus identification with word-topic graph model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1632–1636, Lisbon, Portugal.