SemEval 2022 Task 10: Structured Sentiment Analysis

Jeremy Barnes[1], **Laura Oberländer**[2], **Enrica Troiano**[2], **Andrey Kutuzov**[3]
Jan Buchmann[4], **Rodrigo Agerri**[1], **Lilja Øvrelid**[3], **Erik Velldal**[3]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]IMS, University of Stuttgart
[3]LTG, University of Oslo
[4]Technical University of Darmstadt

{jeremy.barnes, rodrigo.agerri}@ehu.eus
{laura.oberlaender, enrica.troiano}@ims.uni-stuttgart.de
{andreku, liljao, erikve}@ifi.uio.no
buchmann@ukp.informatik.tu-darmstadt.de

Abstract

In this paper, we introduce the first SemEval shared task on *Structured Sentiment Analysis*, for which participants are required to predict all sentiment graphs in a text, where a single sentiment graph is composed of a sentiment holder, target, expression and polarity. This new shared task includes two subtracks (monolingual and cross-lingual) with seven datasets available in five languages, namely Norwegian, Catalan, Basque, Spanish and English. Participants submitted their predictions on a held-out test set and were evaluated on Sentiment Graph F_1. Overall, the task received over 200 submissions from 32 participating teams. We present the results of the 16 teams that provided system descriptions and our own expanded analysis of the test predictions.

1 Introduction

Affective computing is a fundamental step towards enabling human computer interaction (Picard, 1997), as human communication is filled with affective content which conveys a speaker's private state, *i.e.* their current mood, their emotional experiences, or their attitude towards a certain object of conversation. Along with emotion detection, sentiment analysis (Pang et al., 2002; Turney, 2002; Wiebe et al., 2005) is an important stepping stone towards this goal. On a more practical level, being able to automatically determine what people think about an idea, product, or policy is of interest to companies, governments, and private citizens.

In this paper, we describe the SemEval-2022 shared task on *Structured Sentiment Analysis*, which can be thought of as an information extraction problem in which one attempts to find all of the opinion tuples $O = O_i, \ldots, O_n$ in a text. Each opinion O_i is a tuple (h, t, e, p) where h is a **holder** who expresses a **polarity** p towards a **target** t through a **sentiment expression** e, implicitly defining pairwise relationships between elements of the same tuple. Liu (2012) argues that all of these elements are essential to fully resolve the sentiment analysis problem. Although early annotation efforts in sentiment analysis annotated for fine-grained sentiment (Wiebe et al., 2005; Toprak et al., 2010), most research on modeling sentiment *focuses either on a variety of sub-tasks* which avoid performing the full task, *e.g.* targeted (Hu and Liu, 2004), aspect-based (Pontiki et al., 2014), or end-2-end sentiment (Wang et al., 2016), *or instead relies on simplified and idealized tasks*, *e.g.* sentence-level binary polarity classification (Pang et al., 2002).

We argue that the division of fine-grained sentiment into these sub-tasks has become counterproductive, as reported experiments are often not sensitive to whether a given addition to the pipeline improves the overall resolution of sentiment, nor do they take into account the inter-dependencies of the various sub-tasks.

Motivated by this, we present the SemEval-2022 shared task on **Structured Sentiment Analysis**, which jointly predicts all elements of an opinion tuple and their relations.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1280 - 1295

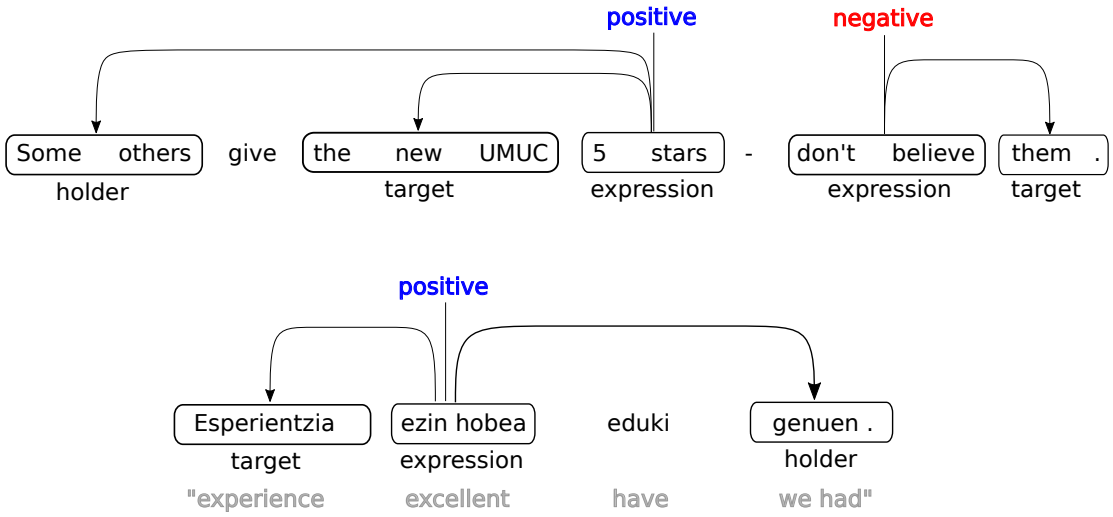

Figure 1: A structured sentiment graph (shown in English and Basque) is composed of a holder, target, sentiment expression, their relationships and a polarity attribute. Holders and targets can be null.

2 Related Work

The conceptual roots of Structured Sentiment Analysis can be found in early computational work on sentiment (Hu and Liu, 2004; Wiebe et al., 2005). Much research in the field has been motivated by the corpus compiled by Wiebe et al. (2005), who annotated English news wire documents with sentiment holders, targets, expressions, intensities, and other variables of interest. Subsets of these variables have been detected with linear (Choi et al., 2006; Yang and Cardie, 2012) and neural models (Katiyar and Cardie, 2016; Zhang et al., 2019), but the full task has never been performed simultaneously.

For instance, various SemEval shared tasks on *Aspect-Based Sentiment Analysis* (ABSA) (Pontiki et al., 2014, 2015, 2016) have focused on target extraction and polarity classification, and there have been research efforts to predict sentiment expressions as well (Wang et al., 2017). The models implemented for that purpose, however, neither resolve relations between expressions nor predict their polarity. The combination of targets, expressions and their polarity have been addressed for the recent task of *aspect sentiment triplet extraction* (Peng et al., 2019; Xu et al., 2020), but the resources that are used for this goal, which typically augment existing targeted datasets with polar expressions, suffer from a major limitation: they do not report annotation guidelines, procedures, or inter-annotator agreement, leaving the final quality of the data unclear.

To solve these issues and integrate all such perspectives, Barnes et al. (2021) proposed a holistic approach to sentiment. They cast the problem of structured sentiment as one of dependency parsing, they introduced specific metrics to evaluate automatic performance on this task, and developed a state-of-the-art structure-aware model. Further improvements were reported by Peng et al. (2021), who proposed a sparse fuzzy attention mechanism to deal with the sparseness of dependency arcs in the dependency models.

3 Task Description

The aim of this shared task is to predict all sentiment graphs in a text (see Figure 1), where a graph includes the elements of an opinion tuple (h, t, e, p). We proposed two subtasks, corresponding to monolingual structured sentiment and cross-lingual structured sentiment. Participants were free to participate in one or both setups, and they had the opportunity to submit a single run on each dataset.

Subtask 1: Monolingual structured sentiment. Models implemented for the first setup had to be trained and tested on the same language. We did not include a closed track, but we asked participants to detail all data used to train models and to make their training reproducible. This also allowed the teams to train multi-lingual models on all of the available training data for structured sentiment – a choice that was made by some of them.

Subtask 2: Cross-lingual structured sentiment. The second task required participants to train on

other languages' train set and test on Spanish, Catalan, and Basque. Again, we allowed any further resource besides the train/dev sets for the test language provided with the shared task.

4 Data

The task contained seven datasets in five languages: Basque (**MultiBooked**$_{\text{EU}}$), Catalan (**MultiBooked**$_{\text{CA}}$) (Barnes et al., 2018), English, with data from **OpeNER**$_{EN}$ (Agerri et al., 2013), Multi-Perspective Question Answer corpus (Wiebe et al., 2005, **MPQA**) and Darmstadt Universities corpus (Toprak et al., 2010, **DS**$_{\text{Unis}}$), Norwegian (Øvrelid et al., 2020, **NoReC**$_{\text{Fine}}$) and Spanish (**OpeNER**$_{ES}$) (Agerri et al., 2013). We directly distributed the datasets in json format, with the exception of the last two corpora, as they have more restrictive licensing. For these, we provided links to the data such that participants could acknowledge their respective terms of use and include scripts to preprocess the data in a uniform manner. Table 1 provides an overview of the datasets and their sentiment annotations.

MultiBooked is a collection of hotel reviews in Basque and Catalan, written by users and collected from booking.com. The data was annotated for structured sentiment with polar expressions, targets, and holder labels, as well as polarity and intensity. Each dataset contains around 1,500 sentences, making them the smallest datasets in the shared task. However, these sentences are more densely annotated than some of the larger corpora. For guidelines and inter-annotator agreement, see Barnes et al. (2018).

OpeNER contains an opinion mining corpus of hotel reviews for six languages (de, en, es, fr, it, nl)[1]. For the purposes of the shared task, we only used the English and Spanish data (Agerri et al., 2013). The reviews were extracted from different booking sites from November 2012 to November 2013. Data collection was designed to ensure that different ratings and languages were included for a given hotel review.

The annotations regard opinion expressions, their respective holders and targets, their polarity and opinion strength. The guidelines were based on the work by Wiebe et al. (2005), which defines an opinion expression as a word (or combination of words) that expresses an attitude of the opinion holder towards a target. The corpus also specifies the relations between holders, targets and opinion expressions in opinion triplets[2].

MPQA annotates English news wire texts with a complex set of annotation types, *i.e.* *agent*, *expressive-subjectivity*, *direct-subjective*, *objective-speech-event*, *attitude*, and *target*. These types are also associated to a number of features and relations between one another. For the purpose of the shared task, we keep only the labels of *agent*, *target*, *direct-subjective*, as well as the *polarity* feature of direct-subjective, which respectively map to our holder, target, polar expression and polarity. We further normalize the polarities such that we have only *positive*, *negative*, and *neutral*. This is the second largest dataset, with a large number of holders, but relatively fewer targets and expressions.

DS$_{\text{Unis}}$ was initially published as part of the Darmstadt Service Review Corpus (DSRC) (Toprak et al., 2010). The DSRC contains reviews of online universities and services annotated with sentiment on the sentence level and fine-grained sentiment on the expression level. The **DS**$_{\text{Unis}}$ data that is part of the task contains only the university reviews, and discards the sentence-level annotations. Toprak et al. (2010) distinguish between polar facts and opinions in their annotation scheme. In order to map the data to the format of the shared task, this distinction is resolved.

NoReC$_{\text{Fine}}$ annotates a subset of the Norwegian Review Corpus (Velldal et al., 2018) for fine-grained sentiment, i.e., including polar expressions, targets and holders, as well as their polarity and intensity. The corpus contains annotations for more than 11k sentences (both subjective sentences and fact-implied ones) taken from professional reviews from a number of different domains, such as screen, music, literature, products and games (Mæhlum et al., 2019). Further details on the annotation procedure, guidelines and inter-annotator agreement can be found in Øvrelid et al. (2020).

5 Evaluation

The main metric for the task is Sentiment Graph F_1 (**SF**$_1$), which attempts to quantify how well a

[1] https://github.com/opener-project

[2] For more information about the annotation guidelines: https://github.com/opener-project/opinion-domain-lexicon-acquisition/blob/master/annotation_guidelines/WP5-guidelinesReviews.pdf

	sentences		holders			targets			expressions			polarity		
	#	avg.	#	avg.	max	#	avg.	max	#	avg.	max	+	neu	−
MultiBooked$_{\text{EU}}$	1,520	10.6	296	1.1	6	1,760	1.4	9	2,319	2.2	10	1,940	0	379
MultiBooked$_{\text{CA}}$	1,676	15.2	237	1.1	7	2,350	2.4	18	2,770	2.6	19	1,743	0	1,027
OpeNER$_{EN}$	2,492	14.8	413	1.0	3	3,843	1.8	21	4,149	2.4	21	2,981	0	1,168
OpeNER$_{ES}$	2,054	17.4	225	1.0	2	3,960	2.2	12	4,386	2.2	15	3,557	0	829
MPQA	10,048	23.3	2,265	2.7	40	2,437	6.3	50	2,794	2.0	14	1,082	465	1,059
DS$_{\text{Unis}}$	2,803	20.0	94	1.2	4	1,601	1.2	6	1,082	1.9	9	612	186	805
NoReC$_{\text{Fine}}$	11,437	16.9	1,128	1.0	12	8,923	2.0	35	11,115	5.0	40	7,547	0	3,557

Table 1: Statistics of the datasets, including number of sentences and average length (in tokens), as well as average and max lengths (in tokens) for holder, target, and expression annotations. Additionally, we include the distribution of polarity – restricted to positive, neutral, and negative – in each dataset.

model captures the full sentiment graph (see Figure 1). For $\mathbf{SF}_1$ each sentiment graph is a tuple of (holder, target, expression, polarity). A true positive is defined as an exact match at graph-level, weighting the overlap between the predicted and gold spans for each element, averaged across all three h, t, e spans. We therefore allow some variability at the token-level (properly weighted), as long as the sentiment graph is predicted.

For precision, we weight the number of correctly predicted tokens divided by the total number of predicted tokens (for recall, we divide instead by the number of gold tokens). Correctly predicted tokens can also consist of empty holders and targets.

6 Baselines

We provided participants with two strong baselines: 1) a dependency graph prediction model, and 2) a sequence-labeling pipeline.

Dependency Graph The first baseline approaches sentiment graph prediction as a dependency graph prediction task, following Barnes et al. (2021). Each sentiment graph is converted to a head-final dependency representation, where we set the final token of the sentiment expression as a root node, the final token in each holder and token span as the head of the span, with all other tokens within that span as dependents. The labels simply denote the type of relation (target/holder) and for sentiment expressions, they additionally encode the polarity. After converting the data, we use a neural graph parsing model (Dozat and Manning, 2018), which learns to score each possible arc to predict the output structure simply as a collection of all positively scored arcs. The base of the network structure is a BiLSTM that creates contextualized representations $c_1, \ldots, c_n = \text{BiLSTM}(w_1, \ldots, w_n)$ where w_i is the concatenation of a word embedding, POS tag embedding, lemma embedding, and character embedding created by a character-based LSTM for the ith token. The contextualized embeddings are then processed by two feedforward neural networks (FNN), creating specialized representations for potential heads and dependents, and the scores for each possible arc-label combination are computed by a final bilinear transformation.

Sequence Labeling We also include a sequence labeling baseline. Specifically, this approach first trains three separate BiLSTM models to extract holders, targets, and expressions, respectively. It then trains a relation prediction model, which uses a BiLSTM with max pooling to create contextualized representations of 1) the full text, 2) the first element (either a holder or target) and 3) the sentiment expression. These three representations are then concatenated and passed to a linear layer followed by a sigmoid function. The training consists of predicting whether two elements have a relationship or not, converting the problem into a binary classification. During inference, the model starts by predicting all sub-elements. Next, it decides if these have a relationship (prediction > 0.5). Finally, the predictions are combined to form full sentiment graphs.

7 Results and Discussion

32 teams participated, with over 200 submissions in total for the evaluation period. The top 10 results for **Subtask 1** are shown in Table 2 (for **Subtask 2** in Table 3). Nearly all teams perform better than the baselines and the performance of the winning teams constitutes the new state of the art.

	NoReC	MultiBooked		OpeNER		MPQA	DS	
Team	NO	CA	EU	ES	EN	EN	EN	Avg.
zhixiaobao	52.9	72.8	73.9	72.2	76.0	44.7	49.4	63.1
MT-speech	52.4	72.8	73.9	74.2	76.3	41.6	48.5	62.8
Hitachi	53.3	70.9	71.5	73.2	75.6	40.2	46.3	61.6
SLPL	50.4	68.1	72.3	73.5	74.7	37.5	41.0	59.6
sixsixsix	48.3	71.1	68.1	68.6	72.7	37.9	37.3	57.7
KE_AI	48.3	71.1	68.1	68.6	72.7	36.4	37.3	57.5
SeqL	48.4	70.4	70.3	69.8	72.5	25.4	42.0	57.0
LyS_ACoruña	46.2	65.3	68.0	69.2	69.8	34.9	41.4	56.4
ECNU_ICA	49.6	68.4	68.6	62.3	67.6	35.1	49.0	56.1
ohhhmygosh	48.7	65.8	65.1	66.9	71.0	26.9	41.6	55.1
graph baseline	27.2	51.6	54.5	49.5	52.1	12.5	20.4	38.3
seq baseline	12.3	33.8	36.5	24.0	32.9	0.02	0.06	19.9

Table 2: Top 10 systems for the monolingual Sub-task 1 according to Sentiment Graph F_1.

Team	ES	CA	EU	Avg.
MT-speech	64.4	64.3	63.2	64.0
SLPL	61.8	56.2	58.4	58.8
Hitachi	62.8	60.7	52.7	58.7
sixsixsix	60.4	59.6	51.2	57.1
SeqL	58.9	59.3	51.6	56.6
ECNU_ICA	55.1	61.5	53.0	56.6
Mirs	61.7	54.4	52.2	56.1
LyS_ACoruña	57.0	55.4	50.9	54.4
OPI	56.4	58.6	44.4	53.1
KE_AI	56.1	55.2	46.3	52.5

Table 3: Top 10 systems for the cross-lingual Sub-task 2 according to Sentiment Graph F_1.

8 Summary of Participating Systems

In this section, we summarize the top three approaches and then further discuss some commonalities among the remaining teams.

8.1 The ZHIXIAOBAO submission

The best performing team on **Subtask 1** formulated the task similar to Barnes et al. (2021), using a dependency graph parsing approach. They deviate from the original in several important ways. First, they use either RoBERTa_Large (Liu et al., 2019) (for the English datasets) or XLM-RoBERTA_-Large (Conneau et al., 2020) (for non-English) as a feature extractor, rather than multilingual BERT base (Devlin et al., 2019) and further fine-tune the parameters of this model, rather than freezing it. Secondly, they introduce a new attention mechanism to help differentiate 'in span' and 'out of span' tokens, which helps dealing with tokens that are not a part of any sentiment span. Finally, they also use suffix masking for tokens which are broken into subtokens when computing the edge scores.

During experimentation, they found that removing the final LSTM layer from the Barnes et al. (2021) model gave improved performance. They also found that the 'in-label' approach was beneficial. Interestingly, they also found that XLM-RoBERTa often performed better than similarly sized monolingual BERT models, e.g. for Basque, Catalan, and Spanish.

8.2 The MT-Speech submission

The second best team in **Subtask 1** and best team in **Subtask 2** similarly used a dependency graph parsing approach with an XLM-RoBERTa_Large backbone. Given the rather small size of some of the datasets, this team proposes several data augmentation strategies which prove to be effective. The first is to exploit the Masked Language Modeling pre-training task of XLM-RoBERTa to augment the training data. They do this by randomly masking a small percentage of the words in a text which lie outside of the sentiment expression. They then sample up to 5 new sentences for training, putting a threshold to remove unlikely versions. Secondly, they further pre-train the language model on the training data, but using the Masked Language Model objective. They also use data from similar datasets in Portuguese and the English

	Rank		General Approach						Language Models				Others		
	mono	cross	graph	seq. label	offset pred.	QA	pipeline	dep. graph	large	FT domain	mono	XLM-mono	multi-task	data aug.	ensemble
zhixiaobao	1	-	✓					✓	✓		✓	✓		✓	
MT-speech	2	1	✓	aux.				✓	✓	✓		✓	✓	✓	
Hitachi	3	3	✓	✓			✓		✓		✓				✓
SLPL	4	2	✓	✓			✓								
LyS_ACoruña	8	8	✓					✓	✓		✓		✓	✓	
ECNU_ICA	9	6			✓		✓		?		✓				
ISCAS	10	-	✓			✓	✓		✓		✓	✓			✓
OPI	11	9		✓			✓		✓		✓	✓			
HITSZ-HLT	12	-			✓		✓		EN		✓				
MaChAmp	13	26		✓			✓		✓				✓		
Amex	14	-	✓									✓			
SenPoi	16	15		✓			✓		✓			✓			
ETMS@IITKGP	18	11			✓			✓	non-EN		✓	✓			
SPDB	21	14	✓											✓	✓
UFRGSent	22	18				✓	✓					✓			
SSN_MLRG1	27	-		✓			✓								

Table 4: Characteristics of submissions that submitted a system description. We show the rank for subtask 1 (mono) and 2 (cross), followed by the general approach, the use of language models, and other characteristics. FT domain = Fine-tune to the domain, mono = monolingual LM, XLM-mono = multi-lingual language model used for monolingual task.

SemEval Laptop dataset which has been automatically augmented with polar expressions. This data is converted to the structured sentiment format. Finally, they include several auxiliary tasks to predict the spans in a sentiment graph as sequence labeling tasks. They include final polarity classification auxiliary task, which they perform on the Catalonia Independence Corpus (Zotova et al., 2020).

8.3 Hitachi

The third team compare a graph prediction model (Graph) and a sequence-to-sequence (Seq2Seq) approach. Specifically, Graph includes a BIO sequence labeler for span extraction, followed by relation prediction with biaffine classifiers (Dozat and Manning, 2018). For the Seq2Seq model, they serialize the tuples and use large pretrained language models to predict these serialized tuples. As the task required providing token offsets, as a post-processing step they predict the text anchors using a word-alignment tool (Jalili Sabet et al., 2020).

Generally, they find that graph prediction performs better than Seq2Seq. In an extensive analysis, they find that Seq2Seq's need for an external alignment system is a hindrance if that information is truly necessary. On OpeNER, however, Graph is clearly better. Both approaches generally perform worse on examples with more opinions, although Graph is slightly more robust. Finally, they find that Graph is much faster to train, as there is no decoder. They conclude, however, that there is not enough evidence to conclusively show that Graph is better than Seq2Seq.

8.4 General Trends

We now discuss some general trends within the submissions and their possible effect on the results. We summarize these trends in Table 4.

General Approach: In general, the teams with the best performance all use graph prediction models. The top two teams maintained the dependency graph approach of Barnes et al. (2021). Several strong submissions successfully used a pipeline approach of a sequence labeling model to extract spans, followed by a relation prediction model. Three teams preferred offset prediction to the BIO sequence labelling, while two teams formulated the task as question answering.

Language Models Nearly all teams used some form of pre-trained language models to create contextualized token representations. One important factor seems to be the size of the language model used (Base or Large), as the top teams generally used XLM-RoBERTa_Large (Conneau et al., 2020). MT-Speech find that fine-tuning this model using the masked language model task on the provided training data improves the performance. The benefits of using a monolingual model seem to be language dependent, as for English or Spanish, many submissions found monolingual models best,

while for Basque or Catalan, they often found that multi-lingual language models performed better.

Others Finally, two of the top teams used a multi-task learning approach to incorporate further information into their models. Several teams further perform data augmentation to minimize the impact of the smaller datasets. This is a particularly beneficial side-effect explicitly annotating all parts of the sentiment graph, as participants were able to make changes in a controlled way, such that the opinion itself was not changed. Three teams further included ensemble approaches with less success than in other tasks.

9 Further Analysis

To gain insight into the systems' performance, we move on to analyzing their mistakes and correct predictions. We start with a quantitative analysis to understand what types of errors appear most often in the submissions, and how they vary according to the different opinion spans (e.g., h/t), settings (monolingual/cross-lingual) and datasets.

Next, we relate the predictions made in the two sub-tasks to some structural properties of the data. This qualitative analysis aims at shedding light on what makes a tuple O_i easy/difficult to find, and whether the observations that apply to a sub-task generalize well to the other.

9.1 Quantitative Analysis.

For the quantitative analysis, we group the different error types defined by Oberländer and Klinger (2020) in the following way[3]: Too early (▬▭, ▬▬), Too late (▭▬, ▭▬), Other (▭▭, ▬▬), Multiple (▭▬), False Positive (▪□▪/▭), and False Negative (▭).

Figure 2 shows relative frequencies for each error type across all 10 winning teams in the monolingual setting (left) and cross-lingual setting (right)[4]. To obtain the relative frequencies, we flatten the annotations for each span type and treat either a continuous stream of annotated or of unannotated tokens as the base of our predictions. Where a prediction exactly aligns with an annotated span, it counts as a true positive, where there is no prediction on an unannotated part, as a true negative. For all other cases we count this as another error type as grouped above. True positives and negatives are not shown in the plot, but are considered when calculating the relative frequency.

We find that, generally, predicting the correct *holder* of an opinion in the monolingual setting is easier than predicting the *target* or the *polar expression*. This is to be expected and explained by the fact that the holders of an opinion are overall shorter in length, reducing the potential of making mistakes. Interestingly, this finding does not fully hold for the cross-lingual setting where there are more False Negatives (▭) and Multiple (▭▬) for *holder* than for *polar expression*. A finding which holds for both the monolingual and the cross-lingual setting is that the error type that occurs the most apart from False Positives (▪□▪/▭) and False Negatives (▭) is Multiple (▭▬). Multiple occurs the most when models predict *polar expressions* for the monolingual setting and *target* for the cross-lingual setting. This is likely because the *polar expression* and *target* spans are typically longer, which gives the models more options to find several predictions within the span. The other type of errors are infrequent, notable being Too early (▬▭, ▬▬) and Too late (▭▬, ▭▬), both error types occurring unsurprisingly mostly for the *polar expression* spans.

Now we compare the box plots across the sub-tasks. Looking at the medians, we see that these are well separated with the median for relative frequencies of False Negatives (▭) and Multiple (▭▬) for all span types being lower for the monolingual than for the cross-lingual setting. The same pattern for False Positives (▪□▪/▭) can be seen, with the exception of the span type *polar expressions*. The inter-quartile ranges are not very similar to each other while looking at pairs of the same type of errors across the two sub-tasks. We see this aspect in the lengths of the boxes, which differ quite a lot, especially notable here are *holders*, for which the boxes are two times in length for the cross-lingual setting. We observe that the error frequencies are generally more spread for the cross-lingual sub-task, for False Negatives (▭) and Multiple (▭▬) holders, and False Positives (▪□▪/▭) targets. For other span types the trend is not as clear. Also, the overall spreads across the settings are slightly greater for the cross-lingual setup for the same span types. However, looking at the overall spreads is perhaps less informative about dispersion of the frequencies of errors than the comparison of box lengths, because of the outliers we see for the mono-

[3]The top bar shows the gold span, while the bottom corresponds to the predicted span.

[4]Error frequencies across the top 10 systems can be found in Appendix (Table 6 and Table 7).

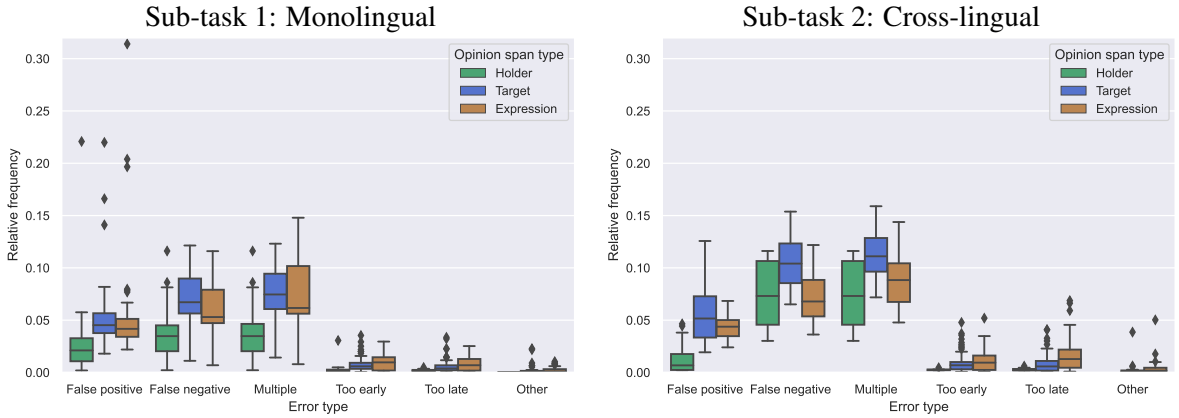

Figure 2: Relative frequencies for each error type for each span type across all teams and datasets.

lingual case. The exceptions are the box plots for the *polar expressions* in the case of the error type Multiple (▄▄). Regarding skewness, we observe generally a right-skewness for the monolingual setup, and no large skew for the cross-lingual setup. For instance, errors on *holders* are skewed to the right in the monolingual case, whereas for the cross-lingual setting, they are slightly skewed to the left. Similarly, *polar expressions* are also skewed to the left for the cross-lingual case and for *targets* the opposite effect can be noticed. Far-away outliers only exist for the monolingual setting, for False Positives (▄▫▄/▄) the far-away outliers are mostly from team `ECNU_ICA` and team `sixsixsix` for the **MPQA** and **NoReC**$_{\text{Fine}}$ datasets. For the False Negatives (▭) and Multiple (▄▄), the far-away outlier is team `ohhhmygosh` for both error types on **MultiBooked**$_{\text{EU}}$ dataset.

The analysis on the box plots evokes a further question: why are there so many False negatives (▭) for *holders* in the cross-lingual setup in comparison to the monolingual one? By inspecting the datasets, we make two observations that explain this result. First, across all datasets used for the cross-lingual sub-task there are only approximately 10% of instances annotated with *holders* and second, the fraction of non-empty *holders* is higher in the test data than in the train data (the numbers can be found in Table 5 in the Appendix).

9.2 Qualitative Analysis.

We now examine the extent to which the correct (or incorrect) identification of specific sub-parts of a sentiment graph (e.g. (h, t)) correlates with another (e.g., (e, p)), or with additional properties of the ground truth (e.g., sentiment intensity). So far, we considered a sentiment graph to be correctly predicted by one system if the system's output was a true positive (following the definition in Section 5). Now, to aggregate the results of all teams, a correctly predicted (ground-truth) graph is one which corresponds to a majority of true positives across teams. Table 8 in Appendix A reports a subsample of texts in which all O_i are predicted correctly (indicated as $+$) by most teams, and those in which most teams did not score a true positive for any of any tuple (marked as $-$).

Within-graph analysis. We begin by focusing on the predictions of h, t, and e. We observe that a successful detection of an opinion holder and target approximates the exact identification of the opinion expression. This is particularly evident in the cross-lingual setup: 75% of ground-truth opinion tuples that have a match in the holder and target for more than half the teams have an exact match in the expression (81% if also the h, t match is exact). In the monolingual setup, this happens in 70% cases (73% for exact h, t span matches).

Typically, if the spans of holder, target and expression are properly recognized, the polarity of such expression is too. It happens for 95% of correctly predicted h, t, e in the cross-lingual setting, for 93% in the monolingual one. These numbers corroborate the relational nature of the span types involved in a structured sentiment task: determining a holder and a target is crucial to establish the opinion linking the two, and hence, its polarity.

Polarity-intensity link. Going beyond the component of a graph, we investigate the relationship between O_i and the intensity of sentiment – a label that was not evaluated in the competition but which characterizes opinions in the used corpora. Figure 3 shows two example distributions of ground-truth

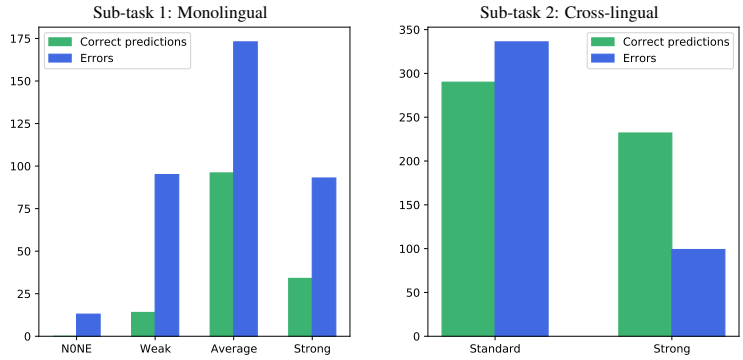

Figure 3: Counts of opinions predicted correctly (with respect to h, t, e, and p) and incorrectly (wrong p): example results from **MPQA** in Sub-task 1, and **OpeNER**$_{ES}$ in Sub-task 2.

opinions for which more than half the teams identified the right (h, t, e, p) (correct predictions), and those where for which (h, t, e) was identified but p was not (errors).

For the setups with only two strength labels, intensity seems to play a role in the prediction of polarity: the submitted systems are consistent in identifying the true polarity of most of the tuples associated to a strong sentiment expression. We see that, for instance in **OpeNER**$_{ES}$, correct predictions are 70% of opinions labelled as "Strong"; the same happens only for 46% of O_i tuples with a standard intensity level. Similar patterns are observed through all corpora in the cross-lingual setup, as well as in the monolingual task based on the same corpora, and **OpeNER**$_{EN}$ (see Appendix Figure 4). This suggests that the recognition of polarity correlates to the intensity with which sentiment is expressed: a sentiment conveyed with higher intensity tends make the prediction "easier".

The link becomes less clear for corpora with non-binary intensity annotation schemas, *i.e.* **MPQA**. Most polarities were not properly recognized across all sentiment strengths. Still, the proportion of wrong-to-correct instances is typically higher with milder intensity degrees (e.g., in **MPQA**, 87% of weak-intensity O_i tuples correspond to errors).

Opinion span sparsity. One feature that differentiates tuples is the sparsity of their spans in the text (e.g., in "*L' habitació un pèl petita per un 5 estrelles*" t and e encompass the whole sentence, while in "*La situació perquè tenim la nostra filla vivint molt a prop . segurament repetirem*", which contains no holder and target, the expression involves only the last two words). Therefore, we observe whether and how the systems' decisions change together with the sparsity of opinion spans, computed as $1 - \frac{\sum \#\text{ tokens}(h,t,e)}{\#\text{ tokens}(text)}$ (Figure 5 in Appendix shows the distribution of sparsity values in the two tasks).

On average, sparsity is lower for correct predictions than for errors. It is 0.73 and 0.72 for the missed h, t, and e spans in the cross-lingual and monolingual tasks, 0.67 (cross-lingual) and 0.66 (monolingual) for the predicted h, t, and e spans. This suggests that spans covering larger parts of the text are easier to predict, while errors tend to occur with labels that are more scattered in the text.

A comparable observation can be drawn relative to the number of opinions in a text. Opinion tuples coming from texts with a higher number of h, t, e relations appear harder to predict: correct predictions occur for tuples that come from texts containing on average 3.4 O_is (in the cross-lingual setup, and 3 in the monolingual task), while errors arise with spans present in text that have on average 4.16 opinions (in the cross-lingual setup, and 3.27 in the other).

10 Conclusion

We proposed to cast sentiment analysis as a structured prediction problem, explicitly predicting the four main elements, and provide seven preprocessed datasets in five languages. Graph prediction models powered by pre-trained language models generally performed best, although several pipeline sequence labelling models also performed well. An analysis of the errors shows that false negatives and predicting shorter spans are the most common errors, while when models correctly predict the holder and target, they generally predict everything correctly. Finally, both more intense polar expressions and spans that cover much of the text are easier to predict.

Acknowledgements

We would like to thank Stephan Oepen for motivation and guidance during the first stages of the shared task. Rodrigo Agerri is funded by the RYC-2017–23647 fellowship and by the ANTIDOTE - EU CHIST-ERA project (PCI2020-120717-2), from the Agencia Estatal de Investigación through the INT-Acciones de Programación Conjunta Internacional (MINECO) 2020 call.

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. In *Procesamiento del Lenguaje Natural*, volume 51, pages 215–218.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney, Australia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Petter Mæhlum, Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, Barcelona, Spain (Online). Association for Computational Linguistics.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2019. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis.

Letian Peng, Zuchao Li, and Hai Zhao. 2021. Sparse fuzzy attention for structured sentiment analysis.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3316–3322. AAAI Press.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-Markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345, Jeju Island, Korea. Association for Computational Linguistics.

Meishan Zhang, Qiansheng Wang, and Guohong Fu. 2019. End-to-end neural opinion extraction with a transition-based model. *Information Systems*, 80:56 – 63.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

A Appendix

A.1 Quantitative Analysis.

A.2 Qualitative Analysis.

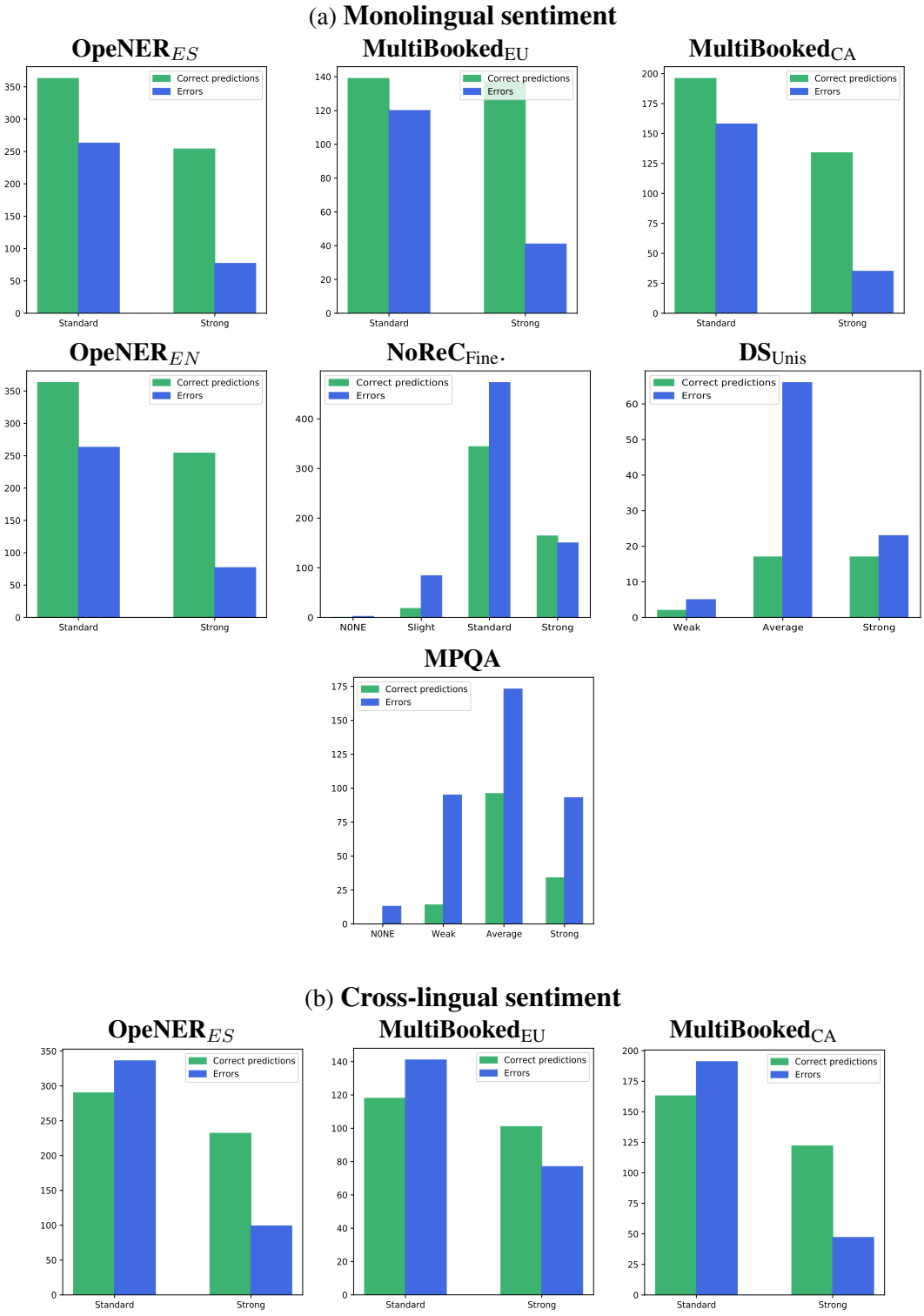

Figure 4: Counts of (h, t, e, p) tuples that are predicted correctly ($>$ half the systems correctly identified all graph components) and incorrectly (wrong p), across intensity levels: (a) Monolingual task, (b) Cross-lingual.

	Holder		Target		Expression	
Dataset	train	test	train	test	train	test
MultiBooked$_{\text{CA}}$	8.50%	10.13%	85.72%	82.98%	100.00%	100.00%
MultiBooked$_{\text{EU}}$	12.21%	13.27%	76.06%	75.74%	100.00%	100.00%
OpeNER$_{ES}$	5.79%	5.85%	90.34%	88.71%	100.00%	100.00%
OpeNER$_{EN}$	9.22%	11.33%	92.89%	91.68%	100.00%	100.00%
DS$_{\text{Unis}}$	7.82%	9.23%	100.00%	100.00%	100.00%	100.00%
NoReC$_{\text{Fine}}$	10.63%	8.91%	80.23%	80.40%	100.00%	100.00%
MPQA	84.06%	83.78%	86.81%	89.19%	100.00%	100.00%

Table 5: The fraction of non-empty annotations for each span type across datasets.

Monolingual task. Cross-lingual task.

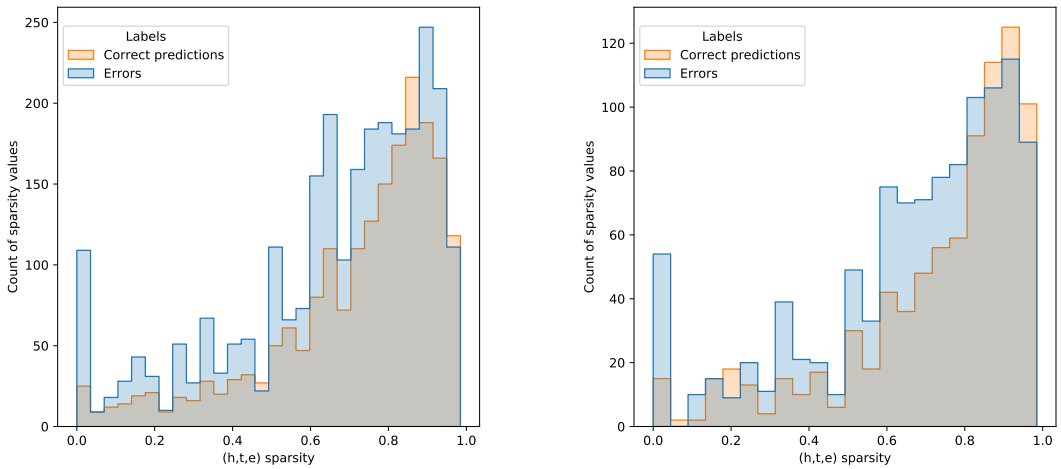

Figure 5: Distribution of sparsity values for h, t, e, spans predicted correctly or incorrectly by most teams in the two setups.

Team	Error type	NoReC			MultiBooked						OpeNER						MPQA			DS		
		NO			CA			EU			EN			ES			EN			EN		
		H	T	E	H	T	E	H	T	E	H	T	E	H	T	E	H	T	E	H	T	E
zhixiaobao	False negative	2	9	8	4	7	5	3	5	4	4	5	5	2	6	5	5	5	5	2	9	9
	False positive	2	5	5	1	4	4	3	6	3	1	4	3	2	4	4	3	3	3	1	7	6
	Multiple	2	10	12	4	7	7	3	5	5	4	6	6	2	7	7	5	5	6	2	9	9
	Other	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Too early	0	1	4	0	2	2	0	2	3	0	1	2	0	2	3	0	1	2	0	1	0
	Too late	0	1	3	0	1	2	0	0	2	0	1	2	0	1	3	0	1	1	0	1	1
Cong666 (MT-speech)	False negative	2	9	8	5	6	4	3	4	4	3	5	4	2	6	5	5	5	5	2	9	8
	False positive	2	5	4	1	5	4	4	6	4	2	4	3	4	4	3	3	3	3	1	8	8
	Multiple	2	10	11	5	7	5	3	5	5	3	5	5	2	7	6	5	6	6	2	9	8
	Other	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Too early	0	1	4	0	2	3	0	1	4	0	2	3	0	2	3	0	1	2	0	1	1
	Too late	0	1	3	0	1	2	0	1	2	0	1	2	0	2	3	0	1	1	0	1	1
Hitachi	False negative	1	9	7	4	6	5	4	5	4	3	5	4	2	5	3	4	4	4	2	10	10
	False positive	2	7	5	1	5	4	3	5	3	2	4	3	3	5	4	4	4	4	0	6	5
	Multiple	1	9	11	4	7	6	4	6	5	3	6	5	2	5	5	5	5	5	2	10	10
	Other	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
	Too early	0	1	4	0	2	3	0	1	3	0	1	2	0	3	3	1	1	2	0	0	1
	Too late	0	1	3	0	1	2	0	1	3	0	1	3	0	2	4	0	1	1	0	1	1
colorful	False negative	2	10	9	5	7	5	4	6	4	4	7	5	2	7	4	5	6	5	2	10	9
	False positive	1	5	4	1	6	5	3	4	4	1	4	2	4	4	3	2	2	3	1	7	8
	Multiple	2	11	14	5	8	6	4	7	6	4	7	6	2	8	6	5	7	6	2	10	10
	Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Too early	0	2	3	0	1	2	0	1	2	0	1	2	0	2	3	0	1	1	0	0	1
	Too late	0	1	3	0	2	2	0	1	2	0	1	2	0	2	3	0	0	1	0	1	1
sixsixsix	False negative	1	9	8	4	6	5	4	6	5	5	6	5	2	5	5	5	5	5	2	11	11
	False positive	3	7	6	2	5	4	4	5	5	1	4	3	3	4	4	4	4	5	0	7	7
	Multiple	1	9	11	4	7	6	4	7	5	5	6	6	2	6	6	5	5	5	2	11	11
	Other	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	Too early	0	1	4	0	2	3	0	2	3	0	2	3	0	3	4	0	1	2	0	0	3
	Too late	0	1	3	0	2	2	0	1	2	0	0	2	0	2	4	0	1	1	0	1	1
KE_AI	False negative	1	9	8	4	6	5	4	6	5	5	6	5	2	5	5	5	5	5	2	11	11
	False positive	3	7	6	2	5	4	4	5	5	1	4	3	3	4	4	4	3	5	0	7	7
	Multiple	1	9	11	4	7	6	4	7	5	5	6	6	2	6	6	5	5	5	2	11	11
	Other	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	Too early	0	1	4	0	2	3	0	2	3	0	2	3	0	3	4	0	1	2	0	0	3
	Too late	0	1	3	0	2	2	0	1	2	0	0	2	0	2	4	0	1	1	0	1	1
SeqL	False negative	2	11	9	4	7	5	4	7	5	4	6	6	2	7	5	-	-	-	2	11	11
	False positive	2	5	4	1	4	4	2	3	3	2	3	3	4	3	4	-	-	-	1	5	5
	Multiple	2	11	12	4	7	6	4	7	5	4	7	6	2	8	6	-	-	-	2	11	11
	Other	0	1	1	0	1	0	0	0	0	0	2	0	0	2	0	-	-	-	0	0	0
	Too early	0	2	5	0	3	4	0	3	4	0	3	3	0	5	4	-	-	-	0	0	1
	Too late	0	1	3	0	3	3	0	4	3	0	3	2	0	4	4	-	-	-	0	0	1
LyS_ACoruña	False negative	2	11	9	4	8	5	4	7	5	5	8	7	2	8	6	5	6	5	2	10	10
	False positive	2	6	5	1	5	6	3	4	4	1	3	3	3	4	3	3	3	4	1	8	8
	Multiple	2	11	14	4	9	7	4	8	6	5	9	8	2	9	7	6	7	6	2	11	10
	Other	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	Too early	0	2	3	0	1	2	0	1	2	0	1	2	0	2	3	0	1	1	0	1	0
	Too late	0	1	3	0	1	2	0	1	3	0	1	2	0	2	3	0	1	1	0	1	0
ECNU_ICA	False negative	0	4	4	-	-	-	2	8	7	4	8	6	3	10	7	1	1	1	0	5	5
	False positive	6	22	20	-	-	-	4	8	5	4	5	5	3	4	4	22	14	31	1	17	20
	Multiple	0	4	7	-	-	-	2	8	8	4	8	7	3	11	8	1	1	1	0	5	5
	Other	0	0	0	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Too early	0	1	4	-	-	-	0	1	3	0	2	3	0	2	3	0	1	1	0	0	2
	Too late	0	1	1	-	-	-	0	1	2	0	1	2	0	1	3	0	1	0	0	1	0
ohhhmygosh	False negative	5	12	12	9	10	10	12	9	11	8	9	9	8	10	10	7	6	8	2	11	11
	False positive	0	5	4	0	5	3	0	4	3	0	4	2	0	4	3	4	3	3	0	6	6
	Multiple	5	12	15	9	10	11	12	9	12	8	10	10	8	10	11	7	7	8	2	11	12
	Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Too early	0	1	3	0	2	2	0	1	2	0	1	2	0	2	3	3	1	1	0	0	0
	Too late	0	1	2	0	1	2	0	1	2	0	1	2	0	2	3	0	1	0	0	0	1

Table 6: Relative frequencies (%) of error types at the span level across the top 10 systems for the monolingual Sub-task 1. H: holder; T: target; E: polar expression.

Team	Error type	OpeNER			MultiBooked					
		ES			CA			EU		
		H	T	E	H	T	E	H	T	E
Cong666 (MT-speech)	False negative	3	12	6	6	7	5	9	7	6
	False positive	4	2	3	1	8	5	0	7	4
	Multiple	3	12	7	6	8	7	9	8	8
	Other	0	0	0	0	0	0	0	0	0
	Too early	0	2	3	0	3	4	0	2	3
	Too late	0	1	5	0	1	2	0	1	4
colorful	False negative	4	13	7	7	11	8	12	12	11
	False positive	2	2	3	0	5	4	0	2	3
	Multiple	4	14	9	7	12	10	12	13	12
	Other	0	0	0	0	0	0	0	0	0
	Too early	0	1	2	0	3	2	0	1	3
	Too late	0	1	5	0	3	4	0	1	4
Hitachi	False negative	4	11	5	7	8	6	12	10	11
	False positive	3	3	4	0	7	5	0	6	3
	Multiple	4	11	6	7	8	8	12	11	13
	Other	0	0	0	0	0	0	0	0	1
	Too early	0	2	3	0	4	3	0	2	2
	Too late	0	2	6	0	2	3	0	1	4
sixsixsix	False negative	3	9	5	7	7	4	11	8	9
	False positive	3	5	4	1	7	7	1	11	6
	Multiple	3	10	6	8	7	5	11	9	10
	Other	0	0	1	0	0	0	0	0	0
	Too early	0	2	3	0	2	5	0	2	3
	Too late	0	3	6	0	3	5	0	2	6
SeqL	False negative	3	12	5	7	8	4	11	10	8
	False positive	5	2	2	0	6	5	0	7	5
	Multiple	3	12	6	7	9	6	11	11	10
	Other	0	4	5	0	1	0	0	1	2
	Too early	0	5	6	0	8	6	0	4	3
	Too late	1	4	8	0	3	3	0	3	8
ECNU_ICA	False negative	5	13	8	-	-	-	7	7	8
	False positive	1	3	4	-	-	-	1	13	5
	Multiple	5	14	9	-	-	-	7	8	10
	Other	0	0	1	-	-	-	0	0	0
	Too early	0	2	3	-	-	-	0	2	4
	Too late	0	2	4	-	-	-	0	1	3
Mirs	False negative	5	13	7	7	10	7	12	11	12
	False positive	2	2	3	1	7	5	0	6	5
	Multiple	5	14	9	7	12	9	12	12	14
	Other	0	0	0	0	0	0	0	0	0
	Too early	0	1	2	0	3	1	0	1	1
	Too late	0	1	5	0	2	3	0	1	4
LyS_ACoruña	False negative	3	15	9	6	9	6	11	10	9
	False positive	4	2	3	1	8	5	1	8	5
	Multiple	3	16	10	6	10	8	11	11	12
	Other	0	0	1	0	0	0	0	0	1
	Too early	0	1	2	0	3	3	0	1	2
	Too late	0	2	5	0	1	3	0	2	5
OPI	False negative	4	14	5	6	10	5	12	14	11
	False positive	2	3	4	0	5	5	0	3	3
	Multiple	4	14	7	6	10	7	12	15	13
	Other	0	0	0	0	0	0	0	0	0
	Too early	0	1	3	0	4	3	0	1	2
	Too late	0	2	5	0	2	3	0	0	4
KE_AI	False negative	6	13	7	8	10	7	11	11	10
	False positive	2	4	4	0	5	5	0	8	4
	Multiple	6	13	8	8	11	8	11	12	11
	Other	0	0	2	0	0	1	0	0	1
	Too early	0	2	4	0	3	4	0	2	4
	Too late	0	3	6	0	5	5	0	1	7

Table 7: Relative frequencies (%) of error types at the span level across the Top 10 systems for the cross-lingual Sub-task 2. (H: holder; T: target; E: expression)

Cross-lingual sentiment		
MultiBooked$_{\text{CA}}$	+	A una habitació , mancaba la teuleta de nit i la persiana estava trencada . hi habia una batidora que no funcionaba .
	−	La ubicació , la decoració i la comoditat de els llits
MultiBooked$_{\text{EU}}$	+	Langileak oso jatorrak , laguntzeko prest eta profesionalak .
	−	Iruzkinak euskaraz egiteko aukera ez izatea .
OpeNER$_{ES}$	+	Muy amables y simpaticos , bastante limpio todo .
	−	Los de recepción te aconsejan un poco sobre donde ir y que linea de metro o de bus coger para ir a los destinos .
Monolingual sentiment		
MultiBooked$_{\text{CA}}$	+	El director de l' hotel molt desagradable .
	−	Tot , ben situat a 10 minuts de la ciutat vella .
MultiBooked$_{\text{EU}}$	+	Harreran zeuden langileen arreta ez zen onena izan .
	−	Bigarren aukera gisan izan bazen ere , zorionekoa izan zen Bergenenen alde hartu genuen aukera .
OpeNER$_{ES}$	+	Ideal para pasar un fin de semana de turismo por la capital . .
	−	Muy bien ubicado para ver el musical
OpeNER$_{EN}$	+	Because of renovation work probably my room was not fully ready .
	−	But the receptionist immediately offered me an improved room with riverside view .
DS$_{\text{Unis}}$	+	Courses are rigorous and challenging .
	−	For the sake of time - I will not comment on quality of education - let's assume it is OK compared to it's competitors .
NoReC$_{\text{Fine}}$	+	Dette er dog lett å tilgi når spillbarheten er så overlegen som den er her .
	−	Og spesielt fabelaktig er det når Claire og Jamie er i selskap på slottet i Versailles i andre episode .
MPQA	+	This announcement was met with unanimous condemnation by the international media .
	−	That is a bitter pill to swallow in a thoroughly non-militaristic society such as ours , where the clash of weapons provokes healthy reactions of repulsion .

Table 8: Examples from different setups and corpora. In the texts marked as $+$, the graphs O_i are predicted correctly by more than half the teams. Items marked as $-$ contain graphs for which the majority of teams missed the correct prediction.