OnForumS: The Shared Task on Online Forum Summarisation at MultiLing'15

Mijail Kabadjov
University of Essex
Colchester, United Kingdom
malexa@essex.ac.uk

Josef Steinberger
University of West Bohemia
Pilsen, Czech Republic
jstein@kiv.zcu.cz

Emma Barker
University of Sheffield
Sheffield, United Kingdom
e.barker@sheffield.ac.uk

Udo Kruschwitz
University of Essex
Colchester, United Kingdom
udo@essex.ac.uk

Massimo Poesio
University of Essex
Colchester, United Kingdom
poesio@essex.ac.uk

ABSTRACT

In this paper we present the Online Forum Summarisation (OnForumS) pilot task at MultiLing'15. OnForumS is a pioneering attempt at encompassing automatic summarisation, argumentation mining and sentiment analysis into one shared task and at bringing crowdsourcing to the evaluation of systems for automatic summarisation and argument structure parsing. It covered two languages, English and Italian. Four research groups, each submitting two runs, participated in the task and these complemented with two baseline system runs were evaluated via crowdsourcing. Performance results are presented and briefly discussed. Being the first of its kind, we believe OnForumS'15 was a successful campaign and hope it will establish itself as a valuable exercise in advancing the state-of-the-art in this new emerging area. Current plans are to organise it again jointly with MultiLing in 2017 and to include more languages.

1. INTRODUCTION

Most major online news publishers, such as The Guardian or Le Monde, publish articles on different topics and encourage reader engagement through the provision of an online comment facility. A given news article can often give rise to thousands of reader comments – some related to specific points within the article, others that are replies to previous comments. The high volume of such user-supplied comments suggests the need for automated methods to summarise this content as it would be otherwise impossible to consume such mass of information in a timely fashion by interested parties, such as journalists, news editors, trend and media monitors to mention but a few. For instance, a reporter working on a given news story may be interested in focusing the follow-ups of his story on the aspects that attracted the most interest or caused the greatest reaction by readers in previously published editions.

The problem of producing a digest of such mass of comments, on the other hand, poses an exciting and novel challenge for the summarisation community and touches on at least three areas of research in Natural Language Processing, as are Automatic Summarisation [7, 9], Argumentation Mining [15, 13, 5] and Sentiment Analysis [16, 23].

The Online Forum Summarisation (OnForumS) pilot task at MultiLing'15[1] is a pioneering attempt at encompassing all three areas into one shared task in order to investigate how the mass of comments found on news providers' web sites can be summarised. We posit that a crucial initial step towards that goal is to determine what comments link to, be that either specific points within the text of the article, the global topic of the article, or comments made by other users. This constitutes a linking task. Furthermore, a set of types or labels for a given link may be articulated to capture phenomena such as agreement (e.g., in favour, against) and sentiment (e.g., positive or negative) with respect to the comment target.

The main contribution of this paper is thus two-fold: firstly, the operationalisation of this labelled linking task into a shared task – to our knowledge the first of its kind – and secondly, casting the evaluation of such task as a crowdsourcing campaign – using crowdsourcing for the evaluation of both summarisation and argument structure has been largely under-explored in previous work.

The remainder of the paper is organised as follows, section §2 describes the shared task and the data set collection and preparation, section §3 gives details of the participating research groups and describes their systems, section §4, discusses results and covers the sampling and evaluation strategy harnessing crowdsourcing, section §5 provides a brief literature survey and finally conclusions are drawn with pointers to future work.

2. ONLINE FORUM SUMMARISATION

The Online Forum Summarisation (OnForumS) is a particular specification of the linking task mentioned in the previous section, in which systems take as input a news article with associated comments and are expected to link each comment sentence to article sentences (which, for simplification, are assumed to be the appropriate units here) or to preceding comments and then to label each link for argument structure $in\_favour, against, impartial$ and sentiment $positive, negative, neutral$.[2] Data for the task is col-

FIRE '15, December 04-06, 2015, Gandhinagar, India

DOI: http://dx.doi.org/10.1145/2838706.2838709

[1] http://multiling.iit.demokritos.gr/pages/view/1516/multiling-2015

[2] The search space for links is defined by the union of Cartesian products of article sentences with comment sentences and comment sentences with comment sentences: $AS \times CS \cup CS \times CS$).

Table 1: OnForumS Corpus (Types a, b, c, d explained in sec. §4.1).

Concept	English	Italian
Number of words	43104	34803
Links validated (via crowdsourcing)	2311	1087
All Links	9635	6193
Unique Links and Labels	6576	4138
Unique Links only	5789	4016
Type d Links	3517	2083
Type c Links	2975	2024
Type b Links	63	20
Type a Links	21	11

lected in two languages English and Italian.[3]

Evaluation of systems output is based on the results of a crowdsourcing exercise, which although widely used in other areas such as Machine Translation [6], Opinion Mining [19] and Word Sense Disambiguation [18], to a much lesser extent it has been employed for evaluating Summarisation and never before in previous MultiLing campaigns. In our case, contributors are asked to judge whether potential links and associated labels are correct for each test article and its comments. The crowdsourcing HIT is defined as a validation task as opposed to annotation, that is, contributors are only asked to validate links and labels produced by systems and are not asked to link or label data themselves. Additionally, due to the high volume of system links only a subset of all the links produced by systems is evaluated by extracting a stratified sample.

2.1 Defining the task

Linking comment sentences to article sentences is a useful step towards summarising the mass of comments. For instance, comment sentences linked to the same article sentence can be seen as forming a "cluster" of sentences on a specific point/topic. Moreover, having labels capturing argument structure and sentiment enables computing statistics within such topic clusters on how many readers are in favour or against the point raised by the article sentence and what is the general 'feeling' about it.

Such clusters of linked sentences are not summaries in themselves, but can be seen as digests of the mass of comments and key points covered in news articles (to an extent resembling the idea of 'capsule overview' put forward in [4]).

2.2 Data

Data for the task was collected in English and Italian. A sample data consisting of one article in English and small set of comments and labelled links result of internal pre-pilots was released early on. The official test data set consisted of ten articles from The Guardian (EN) and six articles from La Repubblica (IT) together with corresponding top fifty comments for each article (see Table 1). The top fifty comments were extracted by sorting all comments in descending order by number of likes and number of replies and choosing the top fifty (note that articles may contain thousands of comments).

3. PARTICIPATING GROUPS

Four research groups participated in the OnForumS shared task, each group submitting two runs. In addition, two baseline system runs were included making a total of ten different system runs.

[3]Sample and test data for the task were released in an XML format pre-tokenised and sentence-split (see http://multiling.iit.demokritos.gr/pages/view/1531/task-onforums-data-and-information).

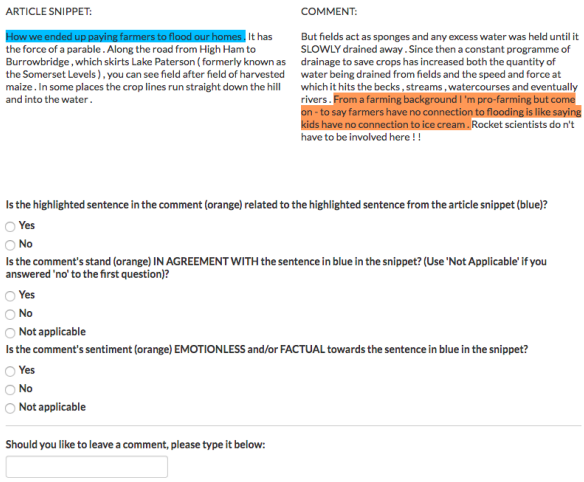


Figure 1: Validation HIT on CrowdFlower.

3.1 Baseline Systems

Our first baseline (FIRST) links every comment sentence to the first sentence of the article which in a news context is typically a good summary of the entire article. In the case of comment responses, it links the response to the first sentences of the parent comment. The second baseline (OVERLAP) links a comment sentence to the article (or parent comment) sentence with the most common words (minimum is 2), stopwords are removed.

Both baselines use the same approach for agreement and sentiment labelling. In the case of sentiment, words from the MPQA lexicon[4] are matched in the comment sentences. If there were more positive than negative words matched, the comment sentence was labeled as positive and vice versa. In the case of a tie, it was classified as neutral. The same approach was taken for argument labelling. In this case, words from Inquirer[5] were used. There are categories Pstv/Affil for words indicating affiliation or supportiveness and Ngtv/Hostile for words indicating an attitude or concern with hostility or aggressiveness. These word counts were used in the same way as discussed for sentiment labelling.

As there are only English versions of the aforementioned dictionaries, we Google-translated them to Italian.

3.2 CIST

The research group at the Center for Intelligence Science and Technology (CIST) of the Beijing University of Posts and Telecommunications (BUPT) divide the task in three parts: content linking, argument labelling and sentiment labelling. They use Word Embedding Model in deep learning combined with WordNet to compute sentence similarity for content linking, and once sentence pairs are linked they implement LDA topic modelling for argument labelling and apply sentiment analysis based on dictionaries for sentiment labelling (see [25] for more details).

For their Word Embedding Model they use GloVe[6] which is a log-bilinear regression model for unsupervised learning of word representations. It needs large amounts of training data which they collect from Wikipedia. Then effectively, every word is mapped into a vector and every sentence into a matrix, whereby a similarity

[4]http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
[5]http://www.wjh.harvard.edu/~inquirer/homecat.htm
[6]http://nlp.stanford.edu/projects/glove

function can be applied (such as cosine distance, for instance).

For argument labelling, they use LDA to discover the latent semantics of sentences with the assumption that documents are represented as a random mixture over latent topics.[7] Then they use the K-means algorithm to cluster sentences into two categories *in_favour* and *against*.

For sentiment labelling, they use three dictionaries as seed, subjectivity, intensifier and valenceshifters lexicons[8], and adopt a machine learning approach to expand them. Then they hand-craft heuristics by experimentation to generate the final sentiment label.

3.3 JRC

The research group at the Joint Research Centre of the European Commission devote most of their effort to the linking and sentiment analysis step and conflate the argument labelling with the sentiment labelling (see [22] for more details).

For sentence linking they employ two methods: a baseline that uses classical lexical based cosine similarity and a more complex method that exploits distributional similarity and term co-occurrence statistics between words.

Their sentiment analysis uses supervised learning with a Support Vector Machines Sequential Minimal Optimization linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists and other features specific for fora and social media.

For the argument labelling, they simply draw on their sentiment labels where *positive* maps to *in_favour* and *negative* to *against*.

3.4 USFD_UNITN

The research groups at the University of Sheffield and the University of Trento develop an elaborate method for linking comments to article sentences prior to labelling sentence pairs for argument and sentiment (see [1] for more details).

They divide the linking step in two phases; firstly they make use of quotes observing that bloggers often quote the segment they are referring to, and secondly, they make use of a richer pre-processing, such as extracting terms and then compute a number of similarity metrics which they combine into one composite similarity score. [9] Finally, a link is created if the similarity score is above a certain threshold determined by experimentation and once a link is created they propagate the link to the whole comment (i.e., all sentences in the comment).

For the argument labelling they train a Support Vector Regressor on a manually annotated corpus, CorEA, using shallow statistical features such as characters, n-grams, punctuation, numbers, etc.

For the sentiment labelling they use GATE[10] which features named entity recognition, event detection, and sentiment detection.

3.5 UWB

The research group at the University of West Bohemia present a system which processes all comment sentences and calculates their similarities to article sentences or parent comment sentences (see [10] for more details). The *similarity score* is based on two models: VSM and LDA. The final score is calculated as an average of similarities computed using both models. At the end of this phase, there is a list of link candidates: either comment sentence to article sentence or comment sentence and comment sentence. Candidates with the anchor sentence shorter than six words are filtered out. The final output of the system consists of one percent of links and two percent of links ranked by the similarity score. For training the VSM and LDA models they used the TREC data. For each detected link, sentiments of both sentences were calculated. The sentiment of the comment was used to fill the sentiment label of the task. Both the comment sentence and the linked sentence sentiments were used to assign the agreement label.

[7] LDA models sentence distribution over topics and topic distribution over words.

[8] The latter two are taken from the system OpinionFinder.

[9] The similarity metrics are combined by assigning weights determined from automatically created training data.

[10] https://gate.ac.uk/

4. EVALUATION VIA CROWDSOURCING

The ten system submissions were evaluated via crowdsourcing[11], which is a commonly used method for evaluating HLT systems [19, 6]. The crowdsourcing HIT was designed as a validation task (as opposed to annotation), where each system-proposed link and labels are presented to a human contributor for their validation with both article sentence and comment sentence placed within context (see Fig. 1). Each system-proposed labelled link is, in fact, a quadruple of the form $< C_i, A_j, arg\_label, sent\_label >$ and then the mapping between this and the HIT is as follows (see Fig. 1):

1. $< C_i, A_j >$ are the comment sentence and the article sentence, respectively, and are fed with their surrounding contexts into the top section of the HIT
2. also the above tuple in itself forms a link and hence is validated via the first 'yes/no' question of the HIT
3. $< arg\_label >$ adopts the following values:"in_favour, impartial, against, not_applicable" which map to "IN AGREEMENT WITH, IMPARTIAL TO, IN DISAGREEMENT WITH, IRRELEVANT TO", respectively, before being fed into the second 'yes/no' question
4. similarly, $< sent\_label >$ is one of "positive, neutral, negative, not_applicable" which map to "EXPRESSING POSITIVE EMOTION, EMOTIONLESS and/or FACTUAL, EXPRESSING NEGATIVE EMOTION, IRRELEVANT TO", respectively, before being fed into the third 'yes/no' question

Both the HIT and the instructions for contributors were translated to English and Italian, thus targeting two distinct groups of native speakers.

4.1 OnForumS Evaluation

The approach used for the OnForumS evaluation is IR-inspired and based on the concept of *pooling* used in TREC [20], where the assumption is that possible links that were not proposed by any system are deemed irrelevant. Then from those links proposed by systems, four categories are formed as follows (see Table 1 for the cumulative distribution of each):

(a) links proposed in 4 or more system runs
(b) links proposed in 3 system runs
(c) links proposed in 2 system runs
(d) links proposed only once

Due to the volume of links proposed by systems, a stratified sample was extracted for evaluation based on the following strategy: all of the **a** and **b** links[12], one third of the **c** links selected at random and one third of the **d** links also selected at random (see Table 1 for numbers of links validated via crowdsourcing).

Once the crowdsourcing exercise was completed, correct and incorrect links were counted first for the linking task only based on

[11] We used CrowdFlower: http://www.crowdflower.com

[12] The popular links (**a** and **b**) were not that many, hence, we chose to include all.

Table 2: System Ranking according to Precision: English.

System-run	Linking	System-run	Argument	System-run	Sentiment
BASE-overlap	93.1	CIST-run2	99.3	CIST-run1	95.1
USFD_UNITN-run2	88.7	CIST-run1	99.1	CIST-run2	93.9
UWB-run1	86.5	UWB-run1	97.5	BASE-overlap	93.8
UWB-run2	86.5	UWB-run2	97.5	BASE-first	93.5
JRC-run1	86.2	BASE-first	92.7	USFD_UNITN-run2	92.6
JRC-run2	83.1	JRC-run2	90.7	JRC-run2	90.3
USFD_UNITN-run1	81.9	USFD_UNITN-run1	89.4	USFD_UNITN-run1	89.8
BASE-first	74.3	JRC-run1	88.9	UWB-run1	88.9
CIST-run2	71.8	BASE-overlap	88.6	UWB-run2	88.9
CIST-run1	70.9	USFD_UNITN-run2	86.2	JRC-run1	87.9

Table 3: System Ranking according to Precision: Italian.

System-run	Linking	System-run	Argument	System-run	Sentiment
BASE-overlap	59.1	CIST-run2	1	CIST-run1	66.7
UWB-run1	25	UWB-run1	1	BASE-overlap	50
USFD_UNITN-run1	20	CIST-run1	77.8	JRC-run1	37.5
JRC-run1	15.2	BASE-first	75	BASE-first	33.3
CIST-run1	8.4	BASE-overlap	69.2	UWB-run1	25
CIST-run2	3.3	JRC-run1	44	CIST-run2	0
BASE-first	1.0	USFD_UNITN-run1	0	USFD_UNITN-run1	0

the aggregated judgements provided by Crowd Flower[13] (i.e., number of 'yes' and 'no' answers from contributors). From those links validated as correct, the correct and incorrect argument and sentiment labels were counted (again, number of 'yes' and 'no' answers). Using these counts precision scores were computed and system runs were then ranked based on these precision scores. For the linking task no system surpassed the baseline algorithm based on overlap followed by USFD_UNITN's runs, and scores were substantially higher for English than for Italian (see Tables 2 and 3).

4.2 Estimating Recall

There are two ways to create gold standard links and labels from the validated data. One is direct validation which entails taking all 'yes' validations of links as gold links and then all labels for argument and sentiment with 'yes' validations as the gold labels for those links. And the other way is by exclusion, if all possible labels for a given link except for one have a 'no' validation then this makes the remaining label a gold label (e.g., if it is not "against", nor "impartial", then it is "in_favour"). With these criteria in mind we created a small gold standard set from which precision, recall and F1 can be computed (see Table 4).[14]

From Table 4 we can see that for top systems recall ranged between $45-70\%$ and precision, $24-25\%$, for the labels *In_Favour* and *Positive*, and precision, $3-5\%$ and around 5% for labels *Against* and *Negative*, respectively. A visualisation of systems performance in terms of precision/recall scatter plots is shown on Figure 2 where it can be quickly seen which are high-recall systems and which high-precision (e.g., for *in_favour*, CIST's system yielded high recall whereas USFD_UNITN's one high precision).

[13]An aggregated judgement is based on multiple judgements using CrowdFlower's "agg" method which returns a single "top" result – AKA the contributor response with the highest confidence (agreement weighted by contributor trust) for every given data point (for more details see: https://success.crowdflower.com/hc/en-us/articles/203527635-CML-and-Instructions-CML-Attribute-Aggregation).

[14]We include P/R/F1 measures only for English as for Italian the number of 'yes' responses was substantially smaller, and hence, the gold set of labels too.

5. RELATED WORK

Producing a digest of the mass of comments found on news providers' web sites with their associated news article content lies at the intersection of three areas of research in Natural Language Processing, as are Automatic Summarisation, Argumentation Mining and Sentiment Analysis. Whilst the former has been an active area of research for decades [12, 4, 7, 9], the latter two are newer areas that have gained much interest in recent years [15, 16].

A good literature survey on Automatic Summarisation evaluation (non-crowdsourcing based) can be found in [11] and on Sentiment Analysis in [2].

Argumentation Mining has gained increased interest in recent years, fuelled by annotated corpora becoming available [14, 24, 21] and work spanning from classification of argumentative propositions in online user comments using SVMs [17], to analysing multilogue in order to classify relations between comments [8] and even using Textual Entailment in identifying agreement relations in discourse fora [5].

Finally, work on mining and analysing online forums has mostly focused on automatic reconstruction of replying structure in discussion threads [26] with the aim of improving areas such as question answering or search [3].

6. CONCLUSION

In this paper we presented the Online Forum Summarisation (OnForumS) pilot task at MultiLing'15. OnForumS is a first attempt at encompassing automatic summarisation, argumentation mining and sentiment analysis into one shared task. It is also a pioneer in bringing crowdsourcing to the evaluation of systems for automatic summarisation and argument structure parsing.

We presented the evaluation strategy followed and the performance results for the participating systems.

We see two key challenges ahead: a more immediate one is to aggregate better the crowdsourcing data by using a probabilistic model of annotation [18], and a longer-term one is to bring in into the task definition higher-level units, such as whole interaction threads.

We plan to organise OnForumS again jointly with MultiLing in

Table 4: Results in terms of precision, recall and F1: English (top scores in bold).

	In_Favour			Against			Positive			Negative		
GroupAndRun	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
BASE-first	7.48	28.27	11.31	2.46	6.01	3.35	24.43	22.99	21.97	1.40	2.28	1.68
BASE-overlap	2.26	35.02	4.18	1.07	19.26	1.90	8.27	39.22	12.76	0.65	9.50	1.22
CIST-run1	67.86	24.49	34.94	0.18	1.03	0.28	45.14	24.35	28.58	2.01	2.27	1.97
CIST-run2	**70.79**	**25.18**	**35.99**	0.18	1.17	0.32	**45.61**	**24.64**	**28.72**	2.01	2.47	2.00
JRC-run1	6.78	34.60	10.78	1.15	8.89	2.00	10.01	29.14	12.77	1.37	6.81	2.24
JRC-run2	9.91	31.11	14.39	0.89	4.60	1.43	12.34	26.57	15.36	1.09	4.70	1.64
USFD_UNITN-run1	0.52	43.89	3.34	**5.44**	**5.15**	**4.39**	13.24	26.86	18.93	**3.00**	**5.83**	**6.21**
USFD_UNITN-run2	0.12	50.00	1.18	1.92	3.97	2.44	7.46	29.19	14.50	1.41	4.64	5.59
UWB-run1	12.91	39.16	17.70	0.06	16.67	0.42	6.69	37.75	11.25	0.00	0.00	0.00
UWB-run2	13.78	21.00	14.97	0.06	8.33	0.42	7.26	18.60	9.28	0.00	0.00	0.00

2017 including more languages; currently there are possibilities for Chinese, Arabic, German and French.

Acknowledgments

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement 610916 âĂŞ SENSEI. Special thanks to Rob Gaizauskas and Mark Hepple for valuable discussions in the early stages of inception of the OnForumS task.

7. REFERENCES

[1] A. Aker, F. Celli, A. Funk, E. Kurtic, M. Hepple, and R. Gaizauskas. Sheffield-Trento System for Sentiment and Argument Structure Enhanced Comment-to-Article Linking in the Online News Domain. http://multiling.iit.demokritos.gr/file/download/1577, 2015. [Online; accessed 06-August-2015].

[2] A. Balahur, E. van der Goot, R. Steinberger, and A. Montoyo, editors. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Baltimore (MD), USA, 2014.

[3] S. Bhatia and P. Mitra. Adopting inference networks for online thread retrieval. In *Proceedings of AAAI*, pages 1300–1305, 2010.

[4] B. Boguraev and C. Kennedy. Salience-based content characterisation of text documents. In I. Mani, editor, *Proceedings of the Workshop on Intelligent and Scalable Text Summarization at the Annual Joint Meeting of the ACL/EACL*, Madrid, 1997.

[5] F. Boltuzic and J. Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore (MD), USA, 2014.

[6] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, volume 1, pages 286âĂŞ–295, 2009.

[7] G. Erkan and D. Radev. LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.

[8] D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, and M. Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore (MD), USA, 2014.

[9] G. Giannakopoulos and V. Karkaletsis. Summary evaluation: Together we stand npower-ed. In *Computational Linguistics and Intelligent Text Processing*, pages 436–450. Springer, 2013.

[10] P. Krejzl, J. Steinberger, T. Hercig, and T. Brychcín. UWB Participation in the Multiling's OnForumS Task. http://multiling.iit.demokritos.gr/file/download/1578, 2015. [Online; accessed 06-August-2015].

[11] A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold-standard. *Computational Linguistics*, 39(2):267–300, 2013.

[12] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[13] M.-F. Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? In P. Majumder, M. Mitra, M. Agrawal, and P. Mehta, editors, *Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation (FIRE'13)*, ACM, New York, NY, USA, 2013.

[14] R. M. Palau and M.-F. Moens. Study on the structure of argumentation in case law. In *Proceedings of the Conference on Legal Knowledge and Information Systems*, pages 11–20, 2008.

[15] R. M. Palau and M.-F. Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011.

[16] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

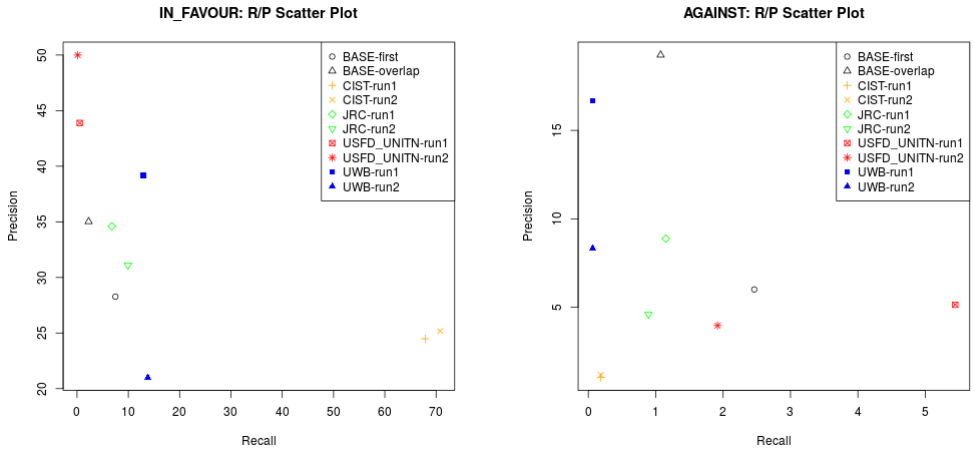
[17] J. Park and C. Cardie. Identifying appropriate support for propositions in on-line user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore (MD), USA, 2014.

[18] R. J. Passonneau and B. Carpenter. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 187âĂŞ–195, Sofia, Bulgaria, August 2013.

[19] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast – but is it good?: Evaluating nonexpert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, pages 254âĂŞ–263, 2008.

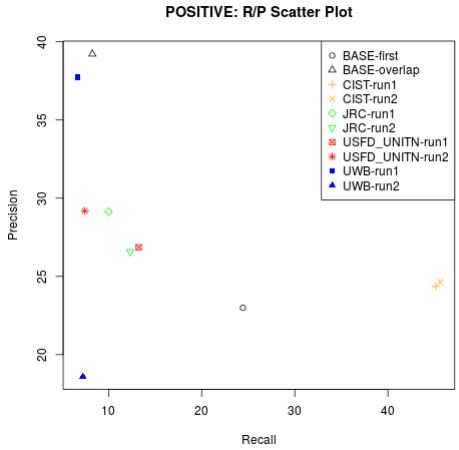
[20] I. Soboroff. Test collection diagnosis and treatment. In *Proceedings of the Third International Workshop on Evaluating Information Access (EVIA)*, pages 34–41, Tokyo, Japan, June 2010.

[21] C. Stab and I. Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING*, pages 1501–1510, 2014.

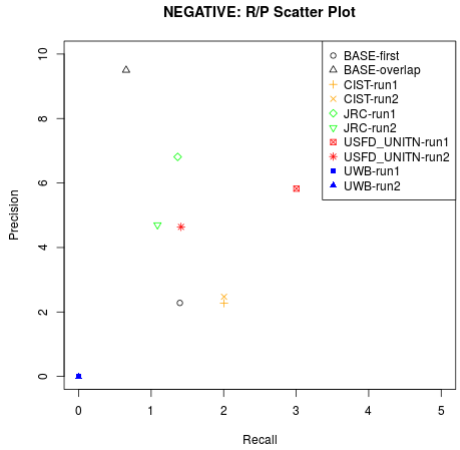


(a) P/R In Favour.

(b) P/R Against



(c) P/R Positive.



(d) P/R Negative

Figure 2: Precision/Recall scatter plots.

[22] H. Tanev and A. Balahur. Tackling the OnForumS Challenge. http://multiling.iit.demokritos.gr/file/download/1576, 2015. [Online; accessed 06-August-2015].

[23] P. Turney and M. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21, 2003.

[24] M. Walker, J. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *Proceedings of LREC*, 2012.

[25] S. Wan, L. Li, T. Huang, Z. Gao, L. Mao, and F. Huang. CIST System Report for SIGdial MultiLing 2015. http://multiling.iit.demokritos.gr/file/download/1575, 2015. [Online; accessed 06-August-2015].

[26] H. Wang, C. Wang, C. Zhai, and J. Han. Learning online discussion structures by conditional random fields. In *Proceedings of ACM SIGIR*, pages 435–444, 2011.