The OnForumS corpus from the Shared Task on Online Forum Summarisation at MultiLing 2015

Mijail Kabadjov*, Udo Kruschwitz*, Massimo Poesio*, Josef Steinberger†, Marc Poch‡, Hugo Zaragoza‡

*University of Essex, Colchester, UK,
{malexa,udo,poesio}@essex.ac.uk
†University of West Bohemia, Pilsen, Czech Republic,
jstein@kiv.zcu.cz
‡Websays, Barcelona, Spain
{marc.poch,hugo.zaragoza}@websays.com

Abstract

In this paper we present the OnForumS corpus developed for the shared task of the same name on Online Forum Summarisation (OnForumS at MultiLing'15). The corpus consists of a set of news articles with associated readers' comments from The Guardian (English) and La Repubblica (Italian). It comes with four levels of annotation: argument structure, comment-article linking, sentiment and coreference. The former three were produced through crowdsourcing, whereas the latter, by an experienced annotator using a mature annotation scheme. Given its annotation breadth, we believe the corpus will prove a useful resource in stimulating and furthering research in the areas of Argumentation Mining, Summarisation, Sentiment, Coreference and the interlinks therein.

Keywords: Online Forums, Summarization, Argument Structure, Sentiment Analysis

1. Introduction

Internet or online forums are discussion websites where people can hold asynchronous conversations in the form of posted messages. Much work has been devoted in recent years on mining and analysing online forums – as in search (Bhatia and Mitra, 2010; Seo et al., 2009), question answering (Ding et al., 2008; Hong and Davison, 2009), classification of argumentative propositions (Park and Cardie, 2014) and automatic summarisation (Giannakopoulos et al., 2015) – and such work, to a large extent is fuelled by the creation and public release of online forums corpora of various kinds (e.g., (Wang et al., 2011) and the `boards.ie` Forums Dataset as part of the ICWSM 2012 conference[1]).

One type of online forums which is increasingly popular is that of readers' discussions taking place on news publishers sites, such as The Guardian[2] or Le Monde[3]. There is strong interest in such forums, their mining and analysis, by a broad range of information seekers, as are journalists, news editors and trend and media monitors. Yet, to our knowledge currently there is no news-forums corpus easily available, most likely due to copyright restrictions on such data, but also due to its novelty.

Furthermore, the high volume of reader-supplied comments on news-forums suggests the need for automated methods to summarise this content, a problem addressed by the shared task on Online Forum Summarisation (OnForumS) at MultiLing 2015[4] (Kabadjov et al., 2015; Giannakopoulos et al., 2015). In addition to summarisation, the OnForumS task addressed the problem of argument structure identification which is particularly relevant in news forums.

In this paper we present the corpus that emerged as a result from the OnForumS shared task and which is aimed at filling the gap in availability of news-forum corpora. The corpus consists of a set of news articles with associated comments from The Guardian (English), which has a comment moderating system in place, and La Repubblica[5] (Italian). The choice of the sources was mainly driven by the number of comments, both sources attract high volume of comments by readers.

The corpus comes with four levels of annotation: argument structure, comment-article linking, sentiment and coreference. The former three were produced via crowdsourcing, whereas the latter (coreference) was carried out by an experienced annotator.[6]

The remainder of the paper is organised as follows: Section §2. discusses online forums and the OnForumS corpus creation, Section §3. gives details on the argument structure annotation, Section §4. describes the coreference and sentiment annotations and finally conclusions are drawn.

2. Online Forums

Three data sets[7] for mining and analysis of online forums were prepared by (Wang et al., 2011) as part of their work on automatic reconstruction of replying structure in discussion threads.

[1] `http://www.icwsm.org/2012/submitting/datasets/`

[2] `http://www.theguardian.com/`

[3] `http://www.lemonde.fr/`

[4] `http://www.sigdial.org/workshops/conference16/sessions.html`

[5] `http://www.repubblica.it/protagonisti/Vittorio_Zucconi`

[6] The same method for annotating coreference was employed in the project LiveMemories (2008–2011): `http://www.livememories.org/`.

[7] `http://sifaka.cs.uiuc.edu/~wang296/Data/index.html`

Table 1: OnForumS Corpus.

	English	Italian
Number of words	43093	34797
Number of sentences	2054	1619

A data set[8] comprising ten years of discussions from the Irish forum site *boards.ie*[9] was made available at the 6th AAAI ICWSM conference.
As part of the OnForumS shared task we prepared a data set of news articles and associated readers' comments which was enriched through crowdsourcing with argument structure, comment-article linking and sentiment. In the next section we describe its creation.

2.1. OnForumS Corpus Creation

Data for the OnForumS task was collected by Websays[10] in English and Italian.[11] A small sample data set resulting from internal pre-pilots was released early on. The official test data set consisted of ten articles from The Guardian and six articles from La Repubblica 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).
An XML format was specifically designed for the OnForumS task, to store pre-tokenised and sentence-split data preserving comment reply structure. An XML snippet is shown below:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE document SYSTEM "ofs.dtd">
<document id="d283147769">
<articleText>
 ... <s id="s40">... </s>
</articleText>
<commentaries>
 <comment bloggerId="richardbj " id="c0">
  <s id="s41">... </s> ...
  <comment bloggerId="questionandfreedom " id="c1">
   <s id="s49">... </s>
  </comment>
 </comment> ...
</commentaries>
<links>
 <link art_sentence="s114" com_sentence="s105" id="10">
  <argument label="in_favour"/>
  <sentiment label="neutral"/>
 </link> ...
</links>
</document>
```

3. Argument Structure Annotation

3.1. The OnForumS Shared Task

The OnForumS task is a particular specification of a linking task, 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 $positive, negative, neutral$ (for more details see (Kabadjov et al., 2015)).

3.2. OnForumS' view on Argument Structure

Identifying argument structure is currently an active area of research (Palau and Moens, 2011; Stab and Gurevych, 2014). In the context of the OnForumS task, the view of argument structure we adopted was that of articulating a closed set of argument labels for the linking of sentence pairs from readers comments and news articles. On one hand, 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 or topic. On the other hand, having labels capturing argument structure enables computing statistics within such topic clusters on how many readers are in favour or against the point raised by the article sentence. Consider the following example from our corpus:

(1) S_A**:** In September the environment secretary, Owen Paterson, assured us that climate change "is something we can adapt to over time and we are very good as a race at adapting".
$\hookrightarrow C_1$**:** Human adaptability!!!!!!!!!!!!! Tell that to ther first dynasty of Egypt (the ones with the pyramids), who died from hunger due to a 30-year drought, the Minoans (volcanic eruption and tsunami), Babylonians (drought), ...
$\rightarrow C_2$**:** Patronising and cynical comment by the Government. I daresay we can 'adapt' to a certain extent but there are limits.

In example 1, the first comment (C_1) links to article sentence S_A through 'human adaptability' and it expresses a view against the quote given in S_A and then the second comment (C_2) seconds the viewpoint of C_1 (it is actually a reply to C_1).
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 (Boguraev and Kennedy, 1997)).
The argument labels are: *in_favour*, *against*, *neutral* and *not_applicable*. The choice of modelling argument structure with a closed set of labels is a rather pragmatic choice driven, firstly, by the need to capture both argument structure and sentiment whilst modelling these in an integrated manner[12] and, secondly, by the objective to define a feasible shared task cast as a classification problem that can be tackled with standard machine learning algorithms.

3.3. Crowdsourcing Validation

Adopting a more pragmatic view on argument structure also has the advantage that it is suitable for annota-

[8] http://www.icwsm.org/2012/submitting/datasets/
[9] http://www.boards.ie/
[10] https://websays.com/
[11] 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).
[12] The sentiment labels parallel the argument ones and are: *positive*, *negative*, *impartial* and *not_applicable*.

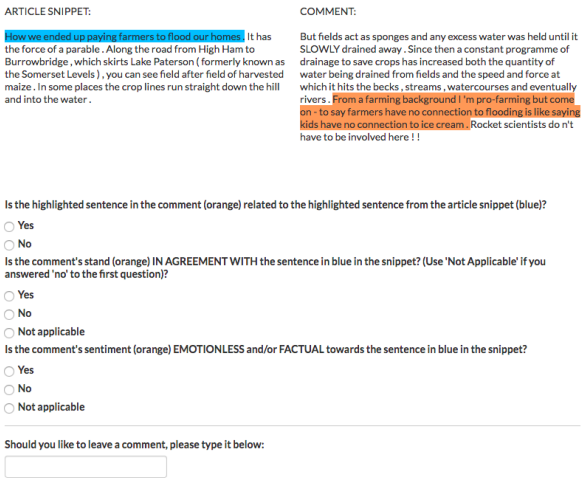

Figure 1: Validation HIT on CrowdFlower.

tion or validation of automatic output using crowdsourcing[13], which is a commonly used method for evaluating Human Language Technology (HLT) systems (Callison-Burch, 2009; Passonneau and Carpenter, 2013). Thus, the crowdsourcing Human Intelligence Task (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).

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

Participation in the crowdsourcing HIT varied between $20-40$ contributors approximately. A sample snapshot of a finished project from CrowdFlower can be seen on Figure 2. For example, the top pane provides information about completeness, cost, number of test questions and rows, whereas the bottom one gives statistics on the contributors, such as number of contributors and contributor satisfaction.

For the crowdsourcing validation we selected a stratified sample from all system proposed links, thus effectively casting any link that was not proposed by any system as irrelevant (akin to the evaluation of IR systems (Soboroff, 2010)). Then from those links proposed by systems the stratification is based on four categories as follows (see Table 2 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

Finally, the sample for validation was selected by taking all links of type **a** and **b** and approximately a third at random from links of type **c** and **d** (see Table 2). For example, for English we took all 21 and 63 **a** and **b** links, respectively, and $2975 * 0.35 \approx 1031$ and $3517 * 0.34 \approx 1196$ of **c** and **d** links, respectively, which makes the set of 2311 links validated (see first row in Table 2).

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

Table 2: OnForumS corpus: link statistics.

	English	Italian
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

Table 3: OnForumS corpus: coreference statistics (TBA: to be annotated).

	English	Italian
Number of markables	14378	*TBA*
Number of coreference chains	1463	*TBA*

From Validation to Annotation. 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.

4. Other Annotations

4.1. Coreference Annotation

The annotation scheme used to annotate for coreference the OnForumS corpus is a variant of the LiveMemories annotation scheme (Rodriguez et al., 2010) which in turn is based on the ARRAU annotation scheme (Poesio and Artstein, 2008). In this corpus all noun phrases are taken as mentions, and the whole noun phrase is considered (with all its embedded NPs). All anaphoric relations of identity between any pairs of mentions are annotated. Coordinations are also treated as mentions, and annotated.

Key coreference statistics are shown in Table 3.

4.2. Sentiment Annotation

As mentioned earlier, the sentiment annotation parallels that of the argument structure annotation and for each comment-article link systems participating in the OnForumS task were supposed to produce a closed set of sentiment labels. These set of labels are: *positive*, *negative*, *impartial* and *not_applicable*. One of the challenges of modelling sentiment is that sentiment is often directed to an entity (i.e., target) which may be mentioned in the article/antecedent sentence or not (e.g., a well-known entity such as 'the government').

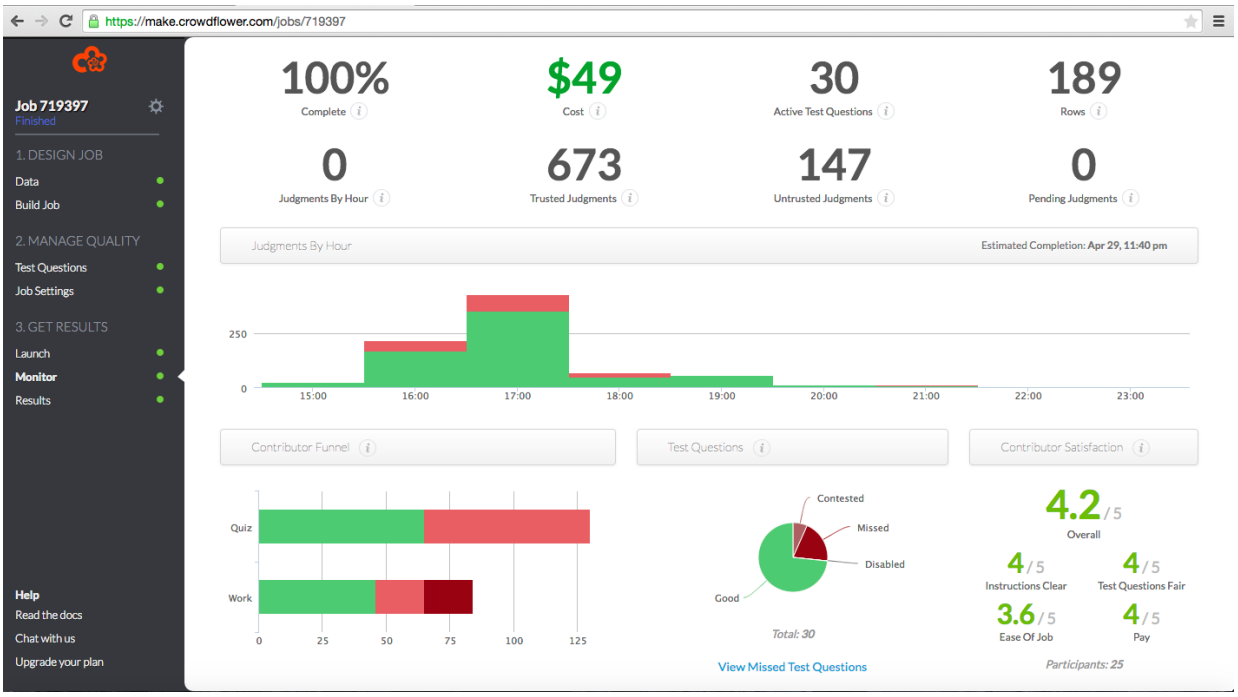

Figure 2: Example of finished CrowdFlower project.

5. Conclusion

In this paper we presented the OnForumS corpus which emerged from the shared task of the same name on Online Forum Summarisation at MultiLing'15. The corpus includes news articles with associated reader's comments from The Guardian (English) and La Repubblica (Italian). It features annotation of argument structure, comment-article linking, sentiment and coreference. The former three were produced through crowdsourcing and came as a result of the evaluation of the official submissions to the OnForumS shared task. The coreference annotation was carried out by an experienced annotator using a mature annotation scheme and methodology for quality assurance.

Given the news-forum domain of the corpus and its annotation breadth, we believe it will prove a useful resource in stimulating and furthering research in the areas of Argumentation Mining, Summarisation, Sentiment, Coreference and the interlinks therein.

Acknowledgments

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement 610916 – SENSEI[14], and project MediaGist, EUs FP7 People Programme (Marie Curie Actions), no 630786. Also, special thanks to Jorge Valderrama and Marco Martinez.

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

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

Callison-Burch, C. (2009). 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 286–295.

Ding, S., Cong, G., Lin, C.-Y., and Zhu, X. (2008). Using conditional random fields to extract contexts and answers of questions from online forums. In *Proceedings of ACL/HLT*, pages 710–718.

Giannakopoulos, G., Kubina, J., Conroy, J. M., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., and Poesio, M. (2015). Multiling 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of SIGdial*, pages 270–274.

Hong, L. and Davison, B. D. (2009). A classification-based approach to question answering in discussion boards. In *Proceedings of SIGIR*, pages 171–178.

Kabadjov, M., Steinberger, J., Barker, E., Kruschwitz, U., and Poesio, M. (2015). OnForumS: The shared task on online forum summarisation at multiling'15. In *Proceedings of the 7th Forum for Information Retrieval Evaluation FIRE*, pages 21–26.

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

Park, J. and Cardie, C. (2014). 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.

Passonneau, R. J. and Carpenter, B. (2013). The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with*

[14] http://www.sensei-conversation.eu/

Discourse, pages 187–195, Sofia, Bulgaria, August.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the arrau corpus. In *Proceedings of LREC*, Marrakesh, Morocco.

Rodriguez, K., Delogu, F., Versley, Y., Stemle, E. W., and Poesio, M. (2010). Anaphoric annotation of wikipedia and blogs in the live memories corpus. In *Proceedings of LREC*, Floriana, Malta.

Seo, J., Croft, W. B., and Smith, D. A. (2009). Online community search using thread structure. In *Proceedings of CIKM*, pages 1907–1910.

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

Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*, pages 46–56, Doha, Qatar.

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