Multilevel Annotation of Agreement and Disagreement in Italian News Blogs

Fabio Celli, Giuseppe Riccardi, Firoj Alam
Dept. of Computer Science,
University of Trento, Italy
{fabio.celli, giuseppe.riccardi, firoj.alam}@unitn.it

Abstract

In this paper, we present a corpus of news blog conversations in Italian annotated with gold standard agreement/disagreement relations at message and sentence levels. This is the first resource of this kind in Italian. From the analysis of ADRs at the two levels emerged that agreement annotated at message level is consistent and generally reflected at sentence level, and that the structure of disagreement is more complex. The manual error analysis revealed that this resource is useful not only for the analysis of argumentation, but also for the detection of irony/sarcasm in online debates. The corpus and annotation tool are available for research purposes on request.

Keywords: Corpora, Agreement/Disagreement, Social Media, News Blogs

1. Introduction and related Work

On-line social conversations concur to the formation of opinions and shared knowledge which influence decision makers. A large amount of multiparty conversations take place online every day in social forums and news blogs (Ruiz et al., 2011), and participants express agreement and disagreement with respect to each others' positions and statements. From a communication analysis perspective, conversation in social media are asynchronous and participants can reply to any other, using text messages or pre-coded actions (e.g. *like* buttons). Previous work on Agreement/Disagreement Relations (henceforth ADRs) in asynchronous online debates, focused either on messages and overall positions of participants (Murakami and Raymond, 2010) (Somasundaran and Wiebe, 2009) (Abu-Jbara et al., 2012), or on the detection of ADRs in pairs of candidate sentences or parts of messages (Andreas et al., 2012). Motivated by the interest in the analysis of argumentation structures in asynchronous conversations, we produced a corpus annotated at message and sentence level in Italian. To do so, we developed a specific annotation tool. Both the corpus and the tool are available for research purposes[1] under a LGPL license. To the best of our knowledge, this is the first resource of ADRs in Italian. We believe that the two levels of annotations maybe useful for argumentation analysis (Schneider et al., 2013) as well as summarization (Di Fabbrizio et al., 2014), irony/sarcasm detection (Reyes et al., 2013) and other kind of parasemantic analyses in the social media domain (Basile and Nissim, 2013), (Celli and Polonio, 2013).

2. Definitions of ADRs

ADRs in conversations can be defined in general terms as shared public commitments, that ground the speech acts performed by the bloggers within the conversations (Lascarides and Asher, 2008). Figure 1 reports an exmple of messages in agreement and disagreement to a news article. From an operational point of view, previous works in asynchronous conversations defined ADRs in different ways. Bender considered ADRs as relationships among bloggers to a multiparty conversation, expressed at message level, with a post or turn text unit (Bender et al., 2011); Walker defined ADRs as Quote-Response message pairs and triplets (chains of three messages such that the third one is a response to the second one which is itself a response to the first one). These pairs and triplets are linked by the structure of the thread, where each message is a reply to its parent and is about the same topic (Misra and Walker, 2013) (Walker et al., 2012) (Morgan et al., 2013). Andreas defined ADRs between pairs of sentences, belonging to messages in a parent/child relation. In their definition, ADRs have a type ("agree", "disagree" or "none") and a mode ("direct" or "indirect", "response" or "paraphrase"). Wang targeted ADRs between text segments corresponding to one or several sentences (Wang and Cardie, 2014). Celli (Celli et al., 2014) defined the ADRs as a function that maps pairs of bloggers and messages to polarity values between 1 ("agree") and -1 ("disagree").

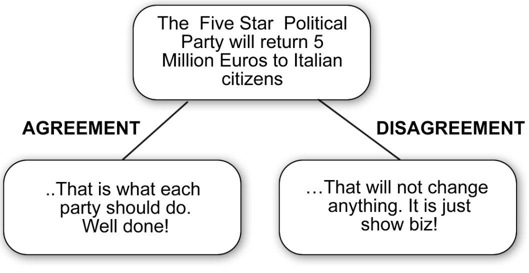
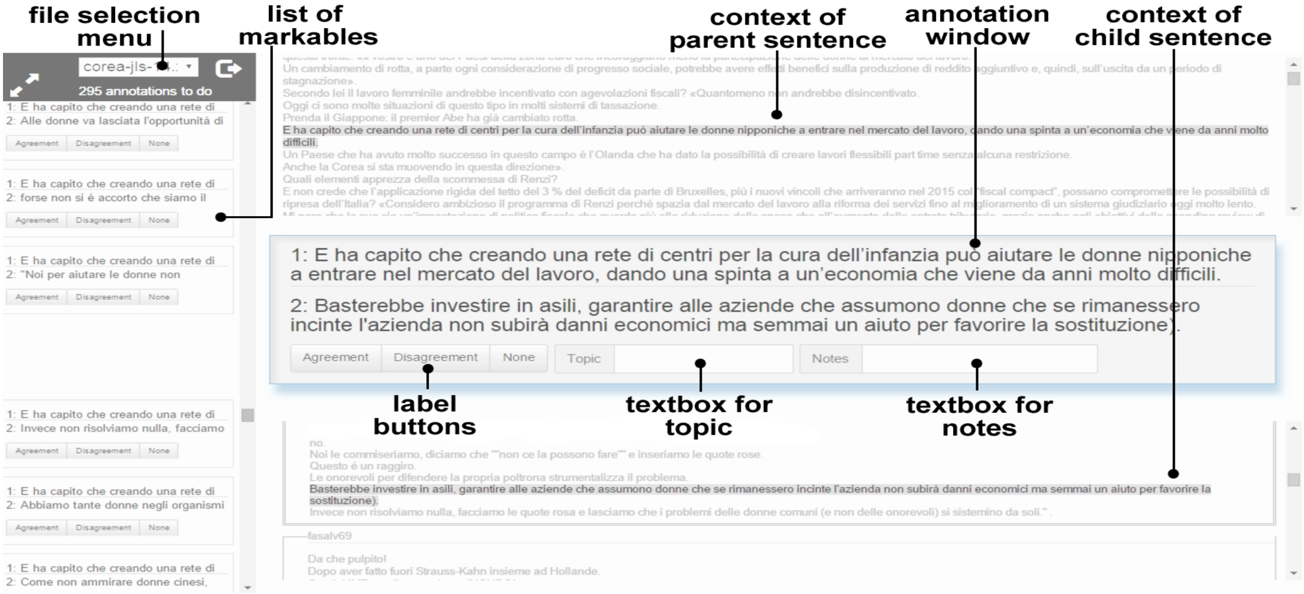

The Five Star Political Party will return 5 Million Euros to Italian citizens
AGREEMENT
DISAGREEMENT
..That is what each party should do. Well done!
...That will not change anything. It is just show biz!

Figure 1: Example of agreement and disagreement relations.

3. Annotation of ADRs

There are few corpora of asynchronous conversations in the social media domain annotated with ADRs, most of them are in English, except AAWD in English, Russian and Chinese, and annotated mainly at one level. Table 1 reports an overview on corpora with ADRs. We started from the CorEA corpus, and we extended the annotation from message to the sentence level.

[1] The language resources can requested at http://sisl.disi.unitn.it/



1: E ha capito che creando una rete di
2: Alle donne va lasciata l'opportunità di

1: E ha capito che creando una rete di
2: forse non si è accorto che siamo il

1: E ha capito che creando una rete di
2: "Noi per aiutare le donne non

1: E ha capito che creando una rete di
2: Invece non risolviamo nulla, facciamo

1: E ha capito che creando una rete di
2: Abbiamo tante donne negli organismi

1: E ha capito che creando una rete di
2: Come non ammirare donne cinesi,

questo fronte: «Il vostro è uno dei Paesi della zona euro che incoraggiano meno la partecipazione delle donne al mercato del lavoro.
Un cambiamento di rotta, a parte ogni considerazione di progresso sociale, potrebbe avere effetti benefici sulla produzione di reddito aggiuntivo e, quindi, sull'uscita da un periodo di stagnazione».
Secondo lei il lavoro femminile andrebbe incentivato con agevolazioni fiscali? «Quantomeno non andrebbe disincentivato.
Oggi ci sono molte situazioni di questo tipo in molti sistemi di tassazione.
Prenda il Giappone: il premier Abe ha già cambiato rotta.
E ha capito che creando una rete di centri per la cura dell'infanzia può aiutare le donne nipponiche a entrare nel mercato del lavoro, dando una spinta a un'economia che viene da anni molto difficili.
Un Paese che ha avuto molto successo in questo campo è l'Olanda che ha dato la possibilità di creare lavori flessibili part time senza alcuna restrizione.
Anche la Corea si sta muovendo in questa direzione».
Quali elementi apprezza della scommessa di Renzi?
E non crede che l'applicazione rigida del tetto del 3 % del deficit da parte di Bruxelles, più i nuovi vincoli che arriveranno nel 2015 col "fiscal compact", possano compromettere le possibilità di ripresa dell'Italia? «Considero ambizioso il programma di Renzi perché spazia dal mercato del lavoro alla riforma dei servizi fino al miglioramento di un sistema giudiziario oggi molto lento.
Mi pare che la sua sia un'impostazione di politica fiscale che guarda più alla riduzione della spesa che all'aumento delle entrate tributarie, grazie anche agli obiettivi della spending review di

1: E ha capito che creando una rete di centri per la cura dell'infanzia può aiutare le donne nipponiche a entrare nel mercato del lavoro, dando una spinta a un'economia che viene da anni molto difficili.

2: Basterebbe investire in asili, garantire alle aziende che assumono donne che se rimanessero incinte l'azienda non subirà danni economici ma semmai un aiuto per favorire la sostituzione).

no.
Noi le commiseriamo, diciamo che ""non ce la possono fare"" e inseriamo le quote rose.
Questo é un raggiro.
Le onorevoli per difendere la propria poltrona strumentalizza il problema.
Basterebbe investire in asili, garantire alle aziende che assumono donne che se rimanessero incinte l'azienda non subirà danni economici ma semmai un aiuto per favorire la sostituzione).
Invece non risolviamo nulla, facciamo le quote rosa e lasciamo che i problemi delle donne comuni (e non delle onorevoli) si sistemino da soli." .

fasalv69

Da che pulpito!
Dopo aver fatto fuori Strauss-Kahn insieme ad Hollande.

Figure 2: Screenshot of the interface of the tool for the annotation of ADRs at sentence level.

corpus	tokens	levels	lang.	IAA
AAWD	325k	msg	multi	k=0.5
IAC	73M	seg	en	α=0.62
LW	-	sent	en	k=0.73
CorEA	135k	msg	it	k=0.87

Table 1: Overview of corpora annotated with agreement/disagreement labels. Corpora are divided by size, levels: message (msg), sentence (sent) or segment (seg), language and Inter Annotator Agreement score (IAA).

ADR Annotation at the message level. CorEA is a corpus of news blogs in Italian, containing conversations of 27 news articles (2887 messages, 135K tokens) of different news categories, including politics, economics, technology, sport and gossips. The average number of messages per conversation is 106.4.

Each parent/child message pair in the corpus has been labelled at message level by two annotators with four labels: agreement (positive or supporting tone towards the parent message), disagreement (negative tone towards parent message), neutral (no opinion or tone expressed towards the parent message), none (if the relation between messages is unclear, i.e. contains only links, or mixed, i.e. contains both agreement and disagreement).

ADR Annotation at the sentence level. The annotation space of ADRs at the sentence level is much larger than at the message space. In order to reduce this space, we put some constraints. We have automatically extracted the sentences from messages with a sentence splitter designed to work for multiple languages (Koehn 2005). Then we extracted candidate ADR sentence pairs that would share a common topic the ADR relation was grounded on. We extracted the topics of articles and conversations with Hierarchical LDA (Teh et al., 2006) (McCallum 2002) and used topic matching to automatically select candidate sentence pairs for the manual annotation, keeping only the root topic of the automatically generated tree. Then we further filtered automatically paired sentences with the following constraints:

1) Sentences must be contained into parent-child reply messages;

2) Sentences must share at least one topic;

3) Sentences must be from different messages of distinct authors.

We designed a tool for reducing the human annotation effort, shown in Figure 2. We display the sentence pairs to be annotated at the centre of the annotation space, as well as the intra-message context above and below (see Figure 2) the annotation space. The fields to be annotated include labels (agree/disagree/none), topic, free text notes. The topic field displays the keywords in common between the two sentences and it can be edited by annotators.

Annotation guidelines. We asked three Italian native speakers to annotate the pair of sentences with three labels (agreement, disagreement or none), topic and notes about the decisions. In the guidelines, we provided the following operational definition for the ADR along with reference examples:

Agreement: sentence B (child) express the same opinion of A (parent) on the same topic or has a positive, supporting tone. E.g. *sentA: I am sure that the boy will make a lot of money from this game! sentB: if he developed the game on his own, for sure he is very smart!*.

Disagreement: sentence B (child) do not express the same opinion on the same topic or has a negative tone towards sentence A (parent). E.g. *sentA: This guy had a great intuition in game design!*. *sentB: I never said the boy is a genius and I never compared him to Steve Jobs, this game is bullshit compared to an OS*.

None: there is no relation between sentence A and B. This case happens in the following conditions: **a) not clear** the annotator cannot understand the relation between sentences (e.g. *sentA: this boy is smart, I think he should take a degree, it is a pity that he does not want to go to the univer-*

sity. sentB: perhaps the boy is lucky); **b) mixed agreement** sentence B contains both agreement and disagreement (e.g. *sentA: this game is awesome! sentB: I played the game, it's funny for the first hour, but then is very boring*); **c) wrong topic**: sentences are not about the same topic (e.g. *sentA: The boy wrote his first program when he was 8 years old. sentB: I think he is not so intelligent, if he does not attend any university program*).

The annotators were asked to commit their ADR label starting from the analysis of the sentence pair and if needed to look into the context to cope with any semantic or discourse ambiguity, but to base their decision on the information grounded in the sentence-pair under evaluation. In the sentence-pair annotation, we removed the neutral class, that proved to be one of the major source of confusion in the annotation at message level.

Evaluation of the Annotation. To evaluate the annotation we compared the IAA between two annotators at sentence and message levels. Results are reported in Table 2. We also evaluated topics at sentence level from annotator's notes: "wrong topic" occurs 13.5% of the times, and 40% of these cases come together with a mismatch between annotators in ADR labeling.

task	examples	classes	*k*
IAA-msg	100	3	0.57
IAA-msg	50	2	0.85
IAA-sent	93	3	0.66
IAA-sent	51	2	0.88

Table 2: inter-annotator agreement (IAA) scores on the annotation of ADRs at message (msg) and sentence (sent) level. The score is computed with Cohen's *k* over 3 and 2 classes.

4. Annotation Analysis and Clustering

At message level, CorEA contains 2887 annotations, at sentence level we produced 5782. The unique messages that contain at least one sentence extracted by topic matching are 1284. By average a message contains 2.34 annotated sentences. Figure 3 shows the distribution of labels at message and sentence level.

The analysis of annotators' notes revealed that 3.2% of the "none" labels at sentence level are due to mixed agreement, 13.5% to wrong topic, as we already mentioned, and 3.1% to unclear cases, when the annotator is not understanding the relation between sentences.

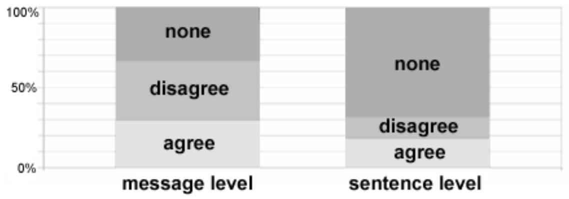


Figure 3: Barplot with the distribution of labels at message and sentence level. We collapsed "none" and "neutral" into "none" at message level.

In order to have an insight on the usefulness of the annotation of ADRs at sentence level for discourse analysis, we measured the purity of labels at the sentence level with respect to the message level. To measure purity, we represented each message as a vector of ADR sentence label counts, and we clustered the vectors with K-means algorithm. With 3 clusters purity is 0.4201, with 4 clusters (represented in Figure 4) is 0.4485. We also ran Expectation maximization clustering to check how many clusters are found automatically: the result is 4 clusters. Clusters are roughly corresponding to Disagreement (cluster1) Agreement (cluster2) and Not Applicable classes (cluster3). Cluster 4 is a collection of noisy examples. This analysis shows that the annotation with 3 classes is generally consistent, however, a small number of noisy examples can be identified.

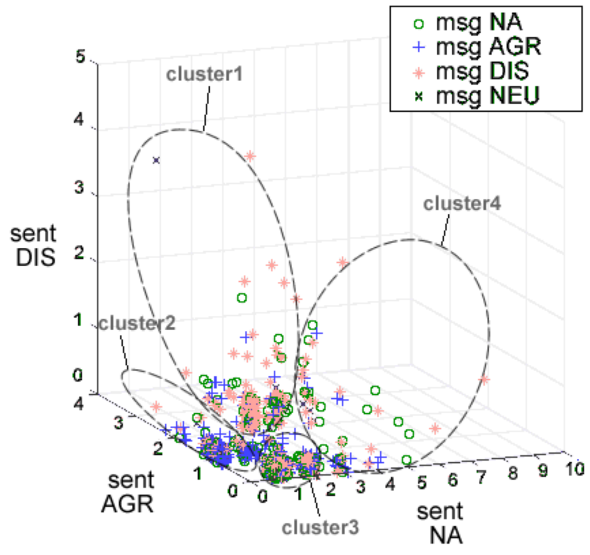


Figure 4: 3D color scatterplot of the distribution of sentence labels with respect to the message labels.

To test the consistency of ADR labels at sentence level, we performed a correlation analysis, that revealed significative correlations (p-value < 0.001) to the message ADR labels, both for Agreement ($\rho = 0.178$) and Disagreement ($\rho = 0.359$).

We also performed a manual error analysis and identified 3 types of problems in the annotation:

1) insufficient information. This occurs when sentences are not informative *per se* out of context. This results in messages annotated either as agreement or disagreement, but with the most sentences in them annotated as "NA". For example these messages are annotated in a Disagreement relation, but there are pairs of sentences in them annotated as "none": *msgA: For example a black orphan shouldn't be adopted by a white family. Nature won't ever let a white couple give birth to a black boy. msgB: Even if you declared yourself not to be homophobic, now you are demonstrating to be racist. You seem to take for granted that only white families can adopt a black boy. Why not viceversa?*; *sentA: Nature won't ever let a white couple give birth to a black boy. sentB: You seem to take for granted that only white families can adopt a black boy.*

2) Sarcasm or irony. This occurs when a message is in a disagree relation with its parent, but sentences in it seem to be in agreement with sentences in its parent: *msgA: Thanks to people like this Italy is becoming poorer and poorer. Trade unions are the real evil in this country! msgB: Ha ha, yes, and mafia, drugs, tax evasion... these are all of secondary importance with respect to the terrible trade unions! sentA: Trade unions are the real evil in this country! sentB: these are all of secondary importance with respect to the terrible trade unions!*
3) Annotators' errors. These cases are sparse and can occur between any class and level.

5. Conclusion

We have annotated a resource for Italian with ADRs at message and sentence levels. This is the first resource of this kind in Italian. From the analysis of ADRs at the two levels emerged that agreement expressed at message level is generally reflected at sentence level, and that the structure of disagreement is more complex. The manual error analysis revealed that this resource is useful not only for the analysis of argumentation, but also for the detection of irony/sarcasm in online debates. The corpus and annotation tool are available for research purposes on request.

6. Acknowledgements

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007–2013) under grant agreement 610916: SENSEI.

7. Bibliographical References

Abu-Jbara, A., Diab, M., Dasigi, P., and Radev, D. (2012). Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 399–409. Association for Computational Linguistics.

Basile, V. and Nissim, M. (2013). Sentiment analysis on italian tweets. *WASSA 2013*, page 100.

Celli, F. and Polonio, L. (2013). Relationships between personality and interactions in facebook. In *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pages 41–54. Nova Science Publishers, Inc.

Di Fabbrizio, G., Stent, A. J., and Gaizauskas, R. (2014). A hybrid approach to multi-document summarization of opinions in reviews. *INLG*, page 54.

Lascarides, A. and Asher, N. (2008). Agreement and disputes in dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 29–36, Columbus, Ohio, June. Association for Computational Linguistics.

Misra, A. and Walker, M. (2013). Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France, August. Association for Computational Linguistics.

Murakami, A. and Raymond, R. (2010). Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Masip, P., and Meso, K. (2011). Public sphere 2.0? the democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics*, pages 1–25.

Schneider, J., Groza, T., and Passant, A. (2013). A review of argumentation for the social semantic web. *Semantic Web*, 4(2):159–218.

Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Wang, L. and Cardie, C. (2014). Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *ACL 2014*, page 97.

8. Language Resource References

Andreas, Jacob and Rosenthal, Sara and McKeown, Kathleen. (2012). Annotating Agreement and Disagreement in Threaded Discussion. In *Proceedings of LREC 2012*. Pages 818–822.

Bender, Emily M and Morgan, Jonathan T and Oxley, Meghan and Zachry, Mark and Hutchinson, Brian and Marin, Alex and Zhang, Bin and Ostendorf, Mari. (2011). Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics. Pages 48–57.

Celli, Fabio and Riccardi, Giuseppe and Ghosh, Arindam. (2014). CorEA: Italian News Corpus with Emotions and Agreement. In *Proceedings of CLIC-it 2014*. Pages 98–102.

McCallum, Andrew K. (2002). MALLET: A Machine Learning for Language Toolkit. *Technical Report*.

Morgan, Jonathan T and Oxley, Meghan and Bender, Emily and Zhu, Liyi and Gracheva, Varya and Zachry, Mark. (2013). Are we there yet?: The development of a corpus annotated for social acts in multilingual online discourse. In *Dialogue & Discourse*, 2(4): 1–33.

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT summit*. Pages 79–86.

Walker, Marilyn A. and Tree, Jean E. and Anand, Pranav and Abbott, Rob and King, Joseph. (2012). A Corpus for Research on Deliberation and Debate. In *Proceedings of LREC*. Pages 812–817.