NewSoMe Corpus of Opinion in News Reports

Corpus Documentation

Basic Information

Corpus name: **NewSoMe Corpus of Opinion in News Reports (NewSoMe News Reports)**

Authors: Roser Saurí (contact Person) – roser.sauri@upf.edu
Judith Domingo – judith.domingom@gmail.com
Toni Badia – toni.badia@upf.edu

Data type: Text

Languages: Catalan (CA), Portuguese (PT), Spanish (ES)

Corpus Size: Data size (uncompressed): 26.5M
Number of words: 945,053

Text Format: Char encoding: UTF-8
Format/Structure: CSV

Corpus Description

The *NewSome Corpus of Opinion in News Reports* is a part of the NewSoMe (News and Social Media) set of corpora, which presents opinion annotations across several genres and covering multiple languages. NewSoMe is the result of an effort to build a unifying annotation framework for analyzing opinion in different genres, ranging from controlled text, such as news reports, to different types of user generated contents: blogs, product reviews, and microblogs (tweets). For further details, Saurí et al. (2014).

In particular, the present corpus covers the genre of news reports and includes about 200 documents for the following languages: Spanish (210 documents), Catalan (224 documents) and Portuguese (200 documents). All the news reports in this corpus have the following annotation layers marked up:

1. **Topic**
   - *Annotation type:* Text extension mark-up.
   - *Description:* Marking up the text segments that express the main topic of the article. The annotation may cover discontinuous elements.
2. **Segment**
   - *Annotation type:* Text extension mark-up.
   - *Description:* Text segment where an opinion is expressed.
3. **Cue**
   - *Annotation type:* Text extension mark-up.
   - *Description:* Text extensions of words and expressions conveying opinion.
4. **Subjectivity**
   - *Annotation type:* classification.
   - *Description:* Classifying the subjectivity nature of any *segment* previously annotated.
   - *Possible values:* Objective, Subjective, NA (Not Applicable).
5. **Polarity**
   - *Annotation type:* Classification.

- *Description:* Classifying the opinion polarity of the news report *topic* and the *segments* identified in a previous annotation phase.
- Possible values:
    - Neutral: No opinion is provided.
    - Polar: Some opinion is provided, but it is unclear whether it is positive or negative (e.g., the author may be using some sort of irony but it is not obvious).
    - Positive: The text expresses a positive opinion.
    - Negative: The text expresses a negative opinion.
    - Mixed: the segment or topic is assessed as both positive and negative.
    - Not Applicable.

6. **Intensity**
    - *Annotation type:* Classification.
    - *Description:* Classifying the polarity strength on the news report *topic* or *segments* previously identified.
    - Possible values: Low, Average, High, NA (Not Applicable).

The annotation has been carried out manually through the crowdsourcing platform CrowdFlower (https://www.mturk.com/), with 7 annotations per annotation layer. The data in the corpus results from aggregating these annotations, and has been used for helping develop an industrial-scale level NLP opinion mining system (Rodríguez-Penagos et al. 2013). The present corpus is a twin of the NewSoMe Corpus for Blogs, also distributed through the Linguistic Data Consortium.

Data Provenance

The documents in the current corpus were crawled from a number of newspaper websites.

For Catalan: www.elperiodico.cat, www.elpunt.cat, www.europapress.es, www.naciodigital.cat, www.diaridegirona.cat, www.lamalla.net, www.radiosio.com, www.radiosio.cat, dbalears.cat, www.vilaweb.cat, www.quelcom.info, www.diariandorra.ad, www.emporda.info, www.elperiodicdandorra.ad, www.ib3.es, www.diaridebalears.com, www.gencat.cat, www.fcbarcelona.com.

For Portuguese: pipocamoderna.mtv.uol.com.br, www2.uol.com.br, www.basketbrasil.com.br, www.monitormercantil.com.br, www.redebomdia.com.br, globoesporte.globo.com, www.ionline.pt, www.campogrande.news.com.br, www.redenoticia.com.br, www.embalagemmarca.com.br, g1.globo.com, www.brasilwiki.com.br, www.estadao.com.br, www.revistafator.com.brwww.orio.pt, www.conjur.com.br, www.midianews.com.br, televisao.uol.com.br, noticias.uol.com.br, entretenimento.uol.com.br, cidadebiz.oi.com.br, cidadebiz.ig.com.br, br.noticias.yahoo.com.

For Spanish: www.abc.es, www.elpais.com, www.elmundo.es, www.20minutos.es.

Corpus Structure and Data Attributes

The Catalan, Portuguese and Spanish parts of the corpus are separated in independent directories: ***news_ca,*** ***news_pt*** and ***news_sp***. Each of them presents the following substructure:

1. ***docs:***
   Directory containing the original documents in text format. It provides crucial information since in the annotation layers based on text extent markup (e.g., *segment, cue*), the information on offset positions is based on these files.
   In the case of the Spanish subcorpus, the annotations at the level of document (*topic,* polarity and *intensity* at the document level) have been carried out on documents different than those annotated at

a finer level than document (*segment, cue,* and *polarity*, *intensity* and *subjectivity* at the *segment* level). Hence, the directory further splits into: *annotated_at_document_level* and *annotated_at_segment_level*.

2. ***sents:***
   Directory with one single file, containing a table with all the sentences in the corpus documents split and identified.
   In the case of the Spanish subcorpus, the file does only contain the data of the documents annotated at a subdocument level (*segment, cue,* and *polarity, intensity* and *subjectivity* at the segment level), given that sentence identification for annotations at the document level were not necessary.

3. ***annots:***
   Directory where annotations are stored. Each annotation layer is in an independent CSV file, listed and described in what follows. Attributes sharing the same name across annotation files encode the same information and can therefore be used for cross-table purposes.
   Annotation layer files:

   a. **topic_annotations.csv**. Marking up the text segments that express the main topic of the article. The annotation may cover discontinuous elements. Table fields:

Attribute	Type	Description
doc_id	ID (key)	Source document identifier.
topic_id	ID (key)	Identifier of the marked-up topic. Format: `tp_INTEGER`.
offset_begin	Integer or NIL	Initial offset of each annotated topic extent in the document.
span_length	Integer	Length of annotated topic extent.
topic_txt	String	Topic textual string.
tag_name	String	Annotation tag name, i.e., `topic`.
agreement	Value	Agreement level among annotators. Possible values: • Expert: Annotation carried out by 1-2 expert annotators. • Plurality: Crowdsourced annotation. Agreement among 2-3 (out of 7) annotators • Majority: Crowdsourced annotation. Agreement among 4-5 annotators. • Absolute: Crowdsourced annotation. Agreement among 6-7 annotators.

   b. **segment_annotations.csv**. Marking up the text segment over which an opinion is expressed. Table fields:

Attribute	Type	Description
doc_id	ID (key)	Source document identifier.
seg_id	ID (key)	Identifier of the marked-up segment relative to its document. Format: `sg_INTEGER`.
offset_begin	Integer	Initial offset of each annotated segment in the document.
span_length	Integer	Length of annotated segment.
seg_txt	String	Segment textual string.
tag_name	String	Annotation tag name, i.e., `segment`.
agreement	Value	Agreement level among annotators. Possible values: • Expert: Annotation carried out by 1-2 expert annotators. • Plurality: Crowdsourced annotation. Agreement among 2-3 (out of 7) annotators • Majority: Crowdsourced annotation. Agreement among 4-5 annotators. • Absolute: Crowdsourced annotation. Agreement among 6-7 annotators.

c. **cue_annotations.csv**. File with cue markups, conveyed via the fields listed below. For each of them, we indicate whether they conform the table key (in DB formatting conventions) and provide its type (ID, integer, string, value).

Attribute	Type	Description
doc_id	ID (key)	Source document identifier.
cue_id	ID (key)	identifier of each annotated cue relative to its document. Format: `qe_INTEGER`.
offset_begin	Integer	Initial offset of each annotated cue in the document.
span_length	Integer	Length of annotated cue.
cue_txt	String	Cue textual string.
tag_name	String	Annotation tag name, i.e., `cue`.
agreement	Value	Agreement level among annotators. Possible values: • Expert: Annotation carried out by 1-2 expert annotators. • Plurality: Crowdsourced annotation. Agreement among 2-3 (out of 7) annotators • Majority: Crowdsourced annotation. Agreement among 4-5 annotators. • Absolute: Crowdsourced annotation. Agreement among 6-7 annotators.

d. **pol_int_doc_annotations.csv**. Indicating the polarity and polarity intensity of each document, hence at the level of *topic* annotations. Table fields:

Attribute	Type	Description
doc_id	ID (key)	Source document identifier.
polarity_val	Value	Polarity value, which can be: Neutral, Polar, Positive, Negative, Mixed, NA (see above).
intensity_val	Value	Intensity value, which can be: Low, Average, High, NA.
polarity_conf	Integer	Degree of confidence on the assigned polarity value.
intensity_conf	String	Degree of confidence on the assigned intensity value.
doc_txt	String	Document text.

e. **pol_int_unit_annotations.csv**. Indicating the polarity and polarity intensity at the level of *segment* markup. Table fields:

Attribute	Type	Description
doc_id	ID (key)	Source document identifier.
sent_id	ID list	Identifying the sentence (or sentences) over which the *segment* markup extends.
seg_id	ID (key)	Identifying the marked-up *segment*.
polarity_val	Value	Polarity value, which can be: Neutral, Polar, Positive, Negative, Mixed, NA (see above).
intensity_val	Value	Intensity value, which can be: Low, Average, High, NA.
polarity_conf	Integer	Degree of confidence on the assigned polarity value.
intensity_conf	String	Degree of confidence on the assigned intensity value.
sent_txt	String	Text of the sentence (or sentences) over which the segment markup extents.

f. **subj_unit_annotations.csv**. Annotating the subjective nature (objective, subjective) of the marked-up segment. Table fields:

Attribute	Type	Description
doc_id	ID (key)	Source document identifier.
sent_id	ID list	Identifying the sentence (or sentences) over which the *segment* markup extends.
seg_id	ID (key)	Identifying the marked-up *segment*.

subjectivity_val	Value	Subjectivity value, which can be: Subjective, Objective, NA.
subjectivity_conf	Integer	Degree of confidence on the assigned polarity value.
sent_txt	String	Text of the sentence (or sentences) over which the segment markup extents.

Acknowledgements

The corpus has been compiled at Barcelona Media Centre d'Innovació (http://www.barcelonamedia.org). Part of the work has been funded by a EU Marie Curie International Reintegration Grant (PIRG04-GA-2008-239414) and by the *Centro para el Desarrollo Tecnológico Industrial (CDTI)*, Spain, as part of the i3Media Project, with the support of the companies: Yahoo! Research Lab Barcelona (http://labs.yahoo.com/Yahho_Research_Barcelona), Acceso (http://www.acceso.com) and Neometrics (http://www.neometrics.com).

References

Rodríguez-Penagos, C., J. Atserias, J. Codina-Filbà, D. García-Narbona, J. Grivolla, P. Lambert, R. Saurí (2013) FBM: Combining lexicon-based ML and heuristics for Social Media Polarities. In: Wilson, T., Z. Kozareva, P. Nakov, S. Rosenthal, V. Stoyanov, and A. Ritter. SemEval-2013 Task 2: Sentiment Analysis in Twitter. Proceedings of the International Workshop on Semantic Evaluation, SemEval. Vol. 13, 2013.

Saurí, R., J. Domingo, T. Badia (2014) The NewSoMe Corpus. A Unifying Opinion Annotation Framework across Genres and in Multiple Languages. Proceedings of the 9th edition of the Language Resources and Evaluation Conference, LREC 2014.