The NewSoMe Corpus: A Unifying Opinion Annotation Framework across Genres and in Multiple Languages

Roser Saurí☼, Judith Domingo♡, Toni Badia☼
☼Department of Translation and Language Sciences, Pompeu Fabra University
♡ Arts and Humanities, Universitat Oberta de Catalunya
roser.sauri@upf.edu, judith.domingom@gmail.com, toni.badia@upf.edu

Abstract

We present the NewSoMe (News and Social Media) Corpus, a set of subcorpora with annotations on opinion expressions across genres (news reports, blogs, product reviews and tweets) and covering multiple languages (English, Spanish, Catalan and Portuguese). NewSoMe is the result of an effort to increase the opinion corpus resources available in languages other than English, and to build a unifying annotation framework for analyzing opinion in different genres, including controlled text, such as news reports, as well as different types of user generated contents (UGC). Given the broad design of the resource, most of the annotation effort was carried out resorting to crowdsourcing platforms: Amazon Mechanical Turk and CrowdFlower. This created an excellent opportunity to research on the feasibility of crowdsourcing methods for annotating big amounts of text in different languages.

Keywords: Corpus creation, opinion mining, crowdsourcing-based annotation.

1. Introduction

Work on opinion mining enjoys a central place within the areas of Natural Language Processing (NLP) and Information Retrieval (IR), mainly due to the direct application of its outcome to the study of market behavior and the design of marketing policies by companies in many sectors. From a technical point of view, opinion mining concerns a variety of tasks that cover different aspects of the opinion expression, such as distinguishing between subjective and objective content, identifying opinion targets and how they are assessed, or detecting the opinion holders.

In order to develop and test opinion mining tools, it is crucial to have consistent and large quantities of corpus annotations covering all these distinctions, but compiling such resources is a well-known bottleneck in the field given the high costs of human annotations, regarding both time and money.

Currently there are a number of opinion corpora available on different genres (mainly for English), which tend to differ in their annotation schemes. Firstly, they vary concerning the element chosen as annotation unit (document, sentence, segment, word). For instance, the MPQA Opinion Corpus of news (Wiebe & Riloff, 2005) offers a fine-grained annotation at the subsentence level, whereas the annotations in the corpus of product reviews by Dave et al. (2003) are applied at the document level.

Secondly, annotation scheme for different genres vary on the entities to tag (e.g., opinion targets, cues, holders of opinion), their properties (classifications such as subjectivity, polarity, strength) or the relations that these hold (for example, between the holder of an opinion and the opinion target). This kind of divergences reflect the natural differences regarding the communicative function and formal properties of each genre. However, there are elements of the opinion expression that are common across genres and thus should be shared by the different annotation schemes.

This paper introduces the **NewSoMe (News and Social Media) Corpus**, a set of subcorpora with annotations on opinion expressions across genres and covering multiple languages. NewSoMe is the result of an effort to: (a) increase the opinion corpus resources available in languages other than English, and (b) build a unifying annotation framework for analyzing opinion in different genres, including controlled text, such as news reports, as well as different types of user generated contents (UGC).

Given the broad design of the resource, and motivated by previous investigations on crowdsourcing corpus annotation (e.g., Mellebeek et al., 2010), most of the annotation effort were carried out resorting to crowdsourcing platforms: Amazon Mechanical Turk (AMT) and CrowdFlower (CF). This opened an excellent opportunity to research on the feasibility of crowdsourcing methods for annotating big amounts of text in different languages.

The paper is structured as follows: the corpus is introduced in section 2., while the annotation effort and the results in terms of inter annotation agreement are presented in sections 3. and 4., respectively. The paper closes with some considerations on the lessons learned from the experience and a brief review of related work.

2. The NewSoMe Corpora

NewSoMe is a set of corpora with texts from a variety of genres and languages containing annotations on opinion expressions. It was developed within an industrial R&D framework on opinion mining, which combined a pressing need for developing industrial-level applications with a natural interest for exploring open questions in NLP research. The practical dimension of this environment required us to develop further opinion corpus resources, either to increase the available corpora for languages and genres already covered, or to cover those not yet contemplated, such as Catalan. At the same time, the large scale of the project appeared as an excellent opportunity to research on the feasibility of crowdsourcing methods for annotating big amounts of text in different languages, along the lines of work introduced in Callison-Burch & Dredze (2010).

2.1. Corpus Design

The NewSoMe project was designed based on the following considerations.

2.1.1. Concerning corpus coverage

Multilinguality. The resource had to expand to languages other than those already covered. For proximity reasons, we chose to include Catalan and Portuguese, while English and Spanish were included given practical considerations in our project (i.e., customer requirements). At the start of the project there were already some resources in these languages, but not for all the targeted genres.

Multiple genres. The resulting resource had to cover a variety of genres in order to be able to offer a broad understanding of how opinions are expressed in language. We chose 4 genres with notable formal and linguistic differences: news reports, web blogs, product reviews, and microblogs (in particular, tweets). There is a gradation concerning their degree of formality and their compliance to the spelling and grammar accepted conventions, ranging from news reports (highly controlled text which follows standardly assumed norms) to the genres of product reviews and tweets. In addition, blogs, product reviews and tweets belong to the super-genre commonly referred to as User Generated Contents (UGC), characterized by a relaxed observance of spelling and grammar conventions, a high use of abbreviations, emoticons, etc., and a strong presence of the author's subjective perspective.

Domain coverage. Product reviews are focused on hotel assessments, thus introducing a domain-oriented component to the NewSoMe corpora. Moreover, blogs divide their contents into the domains of soccer, cooking and miscellany.

Table 1 shows the number of documents for each genre and language included in the NewSoMe set of corpora (EN: English, SP: Spanish, CA: Catalan, PT: Portuguese). In the case of tweets, the document unit corresponds to 1 tweet.

	EN	SP	CA	PT	Total
News reports		200	200	200	**600**
Blogs	108	200			**308**
Product reviews	230	200			**430**
Tweets	1090	8570			**9660**

Table 1: NewSoMe size (in terms of number of documents).

2.1.2. Concerning the annotation scheme

Unifying annotation framework. The communicative function and style specificities of each genre emphasize the use of certain elements of the opinion expressions over others. News reports, for example, encode opinion resorting to expressions of very little (if at all) subjectivity, whereas tweets are loaded with evaluative adjectives, ironical expressions and the use of emphatic markers such as exclamation marks or upper case characters not adhering to the spelling standards.

Differences among genres not only concern grammar and spelling conventions, but also have an impact at the conceptual level. Thus, while opinion targets are always present in product reviews, they are not a necessary element in news reports. The former genre is structured as a set of opinions around a list of different aspects relevant to the main topic of the text (e.g., regarding a phone company: its customer service, the different products it offers, its geographic coverage, etc.). By contrast, news reports tend to articulate opinion as the different views of individual and social actors with regard to their main topic (e.g., the remarks on the Arab Spring by different politicians).

Despite the variety of genres included, we decided that the annotation scheme should apply a unifying view over the data in order to be able to analyze differences in how opinions are expressed across genres, but at the same time be flexible enough so that it can reflect each genre's specificities. Hence, the annotation scheme (presented in section 2.2.) has annotation layers common for aspects that are shared across genres (for example, the distinctions for polarity orientation and its strength), while at the same time presents features exclusive to certain types of text (e.g., the function classification for tweets).

Annotation granularity. There is a divergence concerning the annotation level in previous opinion corpora (document, sentence, sub-sentence), but each of these levels of analysis serves a purpose in the complex task of opinion mining. Thus, opinion annotations should reflect different degrees of granularity: from the broadest scope, at the document level, to expressions with the finest granularity, namely, specific evaluative and subjective words in discourse. We decided that our annotation scheme would encompass different degrees of opinion analysis, from document- to word-level annotations.

2.1.3. Concerning the annotation tool

Crowdsourcing-based annotation. At the starting of the project, the NLP field was immersed in an interesting debate on the feasibility of applying crowdsourcing tools for carrying out time-expensive but crucial NLP tasks, such as developing training corpora or evaluating performance results (Callison-Burch & Dredze, 2010). Mainly, two points were under discussion: whether the data resulting from a crowd of non-experts had a quality level comparable to that in data created by experts, and whether the budget of a crowdsourcing-based project was truly cheaper than that of a comparable project that applied standard methods.[1]

The NewSoMe broad-scope design pointed to a complex and costly process of annotation due to the number of languages and genres to be covered, as well as to the multi-layered annotation approach. At the same time, previous experiments carried out within our research group suggested that opinion annotations could be easily undertaken by non-experts, given that they do not require highly specialized linguistic knowledge but mainly an acceptable understanding of the language and the resources it commonly offers for expressing opinions (Mellebeek et al., 2010).

All these considerations made the project an excellent framework for assessing the feasibility of crowdsourcing

[1] A third issue was introduced into the discussion a bit later, concerning the ethical aspects of resorting to crowdsourcing platforms (Fort et al., 2011). Here we will disregard this consideration.

methods for complex annotations on big amounts of data. We thus set the annotation tasks to be carried out on the crowdsourcing platforms of Amazon Mechanical Turk[2] (AMT) and, at a second stage, CrowdFlower[3] (CF).

2.2. Annotation Scheme

The NewSoMe annotation scheme includes 8 different tags, which aim at covering the most relevant aspects of the opinion expression in the targeted genres. These are:

1. **Document topic** (tag: `topic`), which is the object of opinions expressed at the document level.
2. **Opinion segment** (tag: `segment`), marking up segments of text conveying an opinion.
3. **Opinion target** (tag `target`), for opinions expressed at the sentence level on elements different than the main topic
4. **Cue element** (tag `cue`), representing the words and expressions that convey the opinion on the document `topic` and opinion `targets`.
5. **Text subjectivity** (tag: `subjectivity`). Property applied at the document and target level, with the possible values of: `subjective`, `objective` or `not applicable`.
6. **Opinion polarity** (tag: `polarity`). Property applied at the `topic` and `target` levels, with possible values: `positive`, `negative`, `neutral`, `polar` (for cases that clearly express a polarity but it is not possible to determine what), `mix` (the opinion combines both a positive and a negative assessment), and `not applicable`.
7. **Polarity intensity** (tag: `intensity`), with values: `low`, `average`, and `high`.
8. **Communicative function** (tag: `function`), an attribute specific of tweets and which classifies them as: `target-oriented opinion`, `general opinion`, `sympathetic expression`, `reported fact`, `personal situation`, `other`, and `mixed`. For example, some tweets simply express the whereabouts of its author (e.g., *At the #ACLFest with #wilco*), while others convey an opinion about a target and thus are relevant for opinion annotation (e.g., *I don't think a better band exists to watch as the sun sets. #wilco #ACLFest*). The former tweet is tagged as `personal situation`, whereas the latter corresponds to a `target-based opinion`.

In order to have an exhaustive representation of the opinions expressed in text, these annotations are applied at 2 different levels of granularity: document and subdocument level (which can include segments of text or just single words).

Furthermore, tags in the NewSoMe scheme correspond to three different types:

Text-consuming tags: Tags that mark up text extents in a document. For example, the annotation of `topic`, `target` or `cue`.

Classification tags, such as the judgments on the `subjectivity`, `polarity` and `intensity` at the document level.

Relation-based tags, like the judgments on `polarity` and `intensity` of opinion targets, which connect `target` with `cue` extents.

These tag types are important since they determined to a great extent the customization (and different degrees of success) of the crowdsourcing platforms employed for annotating our data (section 3.).

Tag	Tag Type	Tag Values
`function`	CLA	PS: Personal Situation TO: Target-based Opinion GO: General Opinion SYM: Sympathetic Expression RF: Reported Fact OTH: Other MIX: Mixed
`subjectivity`	CLA	FACT: Factual OPIN: Opinionated NA: Not Applicable
`polarity`	CLA/REL	Non Polar Polar Positive Negative Not Applicable
`intensity`	CLA/REL	Low Average High
`topic`	TEXT	—
`target`	TEXT	—
`cue`	TEXT	—
`segment`	TEXT	—

Table 2: Tag types in NewSoMe. CLA: classification type, REL: relation-based type, TEXT: text-consuming type.

Table 2 shows the type of NewSoMe tags. Classitication tags are accompanied by the list of their possible values. Note that `polarity` and `intensity` can be both classification tags (they classify the polarity and intensity at the document level) and relation-based tags (at the subdocument level, when assessing the polarity on an opinion target according to a given cue).

Some of these levels of analysis are shared among all genres, whereas others are particular to certain types of text. For example, a crucial distinction for tweets concerns their communicative function. Table 3 illustrates common and particular tags for the different genres covered here. Grey areas indicate that the tag is not used in that genre.

The annotation scheme was created combining preliminary data explorations together with the experience in previous related work reported in the literature:

Preliminary empirical explorations. The annotation scheme and its corresponding set of annotation guidelines were established by means of an exhaustive analysis on how opinion is expressed in the different genres contemplated

[2] https://www.mturk.com/

[3] http://crowdflower.com

	function	topic	segment	target	cue	subjectivity	polarity	intensity
News reports								
Document level	▨	✓		▨			✓	✓
Subdocument level	▨		✓	▨	✓	✓	✓	✓
Blogs								
Document level	▨	✓		▨			✓	✓
Subdocument level	▨		✓	▨	✓	✓	✓	✓
Product reviews								
Document level	▨	✓	▨			▨	✓	✓
Subdocument level	▨		▨	✓	✓	▨	✓	✓
Tweets								
Document level	✓	▨	▨			✓		
Subdocument level		▨	▨	✓	✓		✓	✓

Table 3: Tags available at different genres and levels of analysis.

here. In particular, the annotation scheme design was carried out as an iterative process of data analysis, proposal of annotation tags and criteria, testing of the proposal on new data, feedback, and further refinement on the proposal.

For instance, defining the annotation guidelines for tweets involved the participation of two people (with some linguistic formation) who manually annotated 100 tweets randomly chosen. The annotation effort was split based on the different information layers aimed at. Thus, the annotators made separate passes through the corpus for the different tags to be tested (`function`, `subjectivity`, `polarity`, etc.). After the annotation of each layer, a third person acted as a judge validating (or not) the agreements between both annotators, as well as determining the correct annotation in case of disagreement. The judge was the person in charge of designing the scheme and developing the annotation guidelines. Hence, the task of validating agreements and adjudicating disagreement provided her with valuable information concerning how the scheme and annotation guidelines were assimilated by annotators with no previous experience on this annotation tasks or exposition to the data.

Experience from previous work. Together with the empirical analysis of the data, the annotation scheme design is also grounded on the experience attained in previous comparable research. Within NLP, there is a number of projects exploring the same or similar aspects. Many of them, however, focus on opinion as expressed at the document level, thus disregarding the challenges entailed in annotating the specific expressions and components displaying opinion and sentiment in discourse. For instance, many opinion corpora of product reviews present only as their annotated opinions the scores (within a scale of 2, 3, or 5 distinctions) provided by users on the product they review (Pang & Lee, 2008). These are certainly easy-to-compile opinion corpus, but are nevertheless very poor in terms of the degree of elicitation of the information that is necessary for the automatic mining of opinion in text.

Because of that, in the design of the NewSoMe annotation scheme we took into account the experience of previous projects that had explored opinion expression at a fine level of granularity (sentence and subsentence), such as the work on news reports that resulted in the MPQA corpus (Wiebe et al., 2005), or the project on annotating user-generated contents (with particular emphasis on product reviews) reported in Toprak et al. (2010). The former work, for example, was crucial in confirming our conclusions on news reports as a genre for which opinion targets are not obvious to identify, and therefore we followed that experience and decided not to include them in the annotation scheme for news and blogs.

2.3. Data Provenance

The data in this set of corpora was obtained from the following sources:

Tweets: They were collected using the standard Twitter Streaming APIs.[4] Data privacy was ensured by anonymizing author mentions to `@USER` and normalizing hashtags and URLs to the strings `#HASHTAG` and `URL` respectively.

Product reviews: All documents in this corpus are hotel reviews obtained from the online hotel reservation application booking.com (http://www.booking.com).

Blogs: Data was obtained by means of the Google Blog Search API (https://developers.google.com/blog-search/). Spanish blogs belong to the blogging domains of wordpress.com and blogspot.com, while English document were extracted from these same two domains, in addition to asiawrites.org.

News reports: The documents were crawled from a number of newspaper websites.[5]

3. Annotation Effort

Resorting to a crowdsourcing platform for corpus annotation required us to make a number of methodological decisions that need not be considered in annotation projects carried out by means of more standard annotation tools. These decisions concern:

[4] https://dev.twitter.com/docs/streaming-apis

[5] For Catalan: www.elperiodico.cat, www.elpunt.cat, www.europapress.es, dbalears.cat, www.naciodigital.cat, www.vilaweb.cat, www.lamalla.net, www.diaridegirona.cat, www.gencat.cat, www.radiosio.com, www.radiosio.cat, www.quelcom.info, www.ib3.es, www.diariandorra.ad, www.emporda.info, www.elperiodicdandorra.ad, www.diaridebalears.com, 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.campogrande.news.com.br, www.redenoticia.com.br, www.embalagemmarca.com.br, www.brasilwiki.com.br, www.estadao.com.br, www.revistafator.com.br www.orio.pt, www.conjur.com.br, www.midianews.com.br, www.ionline.pt, televisao.uol.com.br, noticias.uol.com.br, g1.globo.com, 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.

1. The annotation tool.
2. Strategies for on-the-fly controlling the quality of the annotations.
3. Methods for aggregating the data annotated by the *turkers* (i.e., the non-expert annotators).

Crowdsourcing platforms facilitate technical resources for this type of actions to customers that want to run their own project. Most of them concern tasks of obtaining contents (e.g., postal addresses in websites) or item classification (be it text, image, audio, etc.). Nevertheless, linguistic annotation also involves marking up text extents or tagging relations among annotated pieces of text, as seen above. The coming subsections detail the methodological and technical solutions we adopted for handling the three issues listed above.

3.1. Customizing the annotation tools

In addition to their ready-to-use basic templates for data gathering, both AMT and CF allow developing your own tools for more complex tasks by combining HTML code (or equivalent) with JavaScript. We resorted to that feature because the three-fold nature of the different tags in the NewSoMe annotation scheme required to use more sophisticated annotation tools than those facilitated through the basic templates.

Annotation tools for text-consuming tags: We developed our own interface for displaying text and allowing the *turkers* to select (or deselect) fragments of text.

Annotation tools for classification tags: Templates for classification tasks are a basic staple in both AMT and CF, and therefore we took advantage of that customizing them to our data and information to annotate.

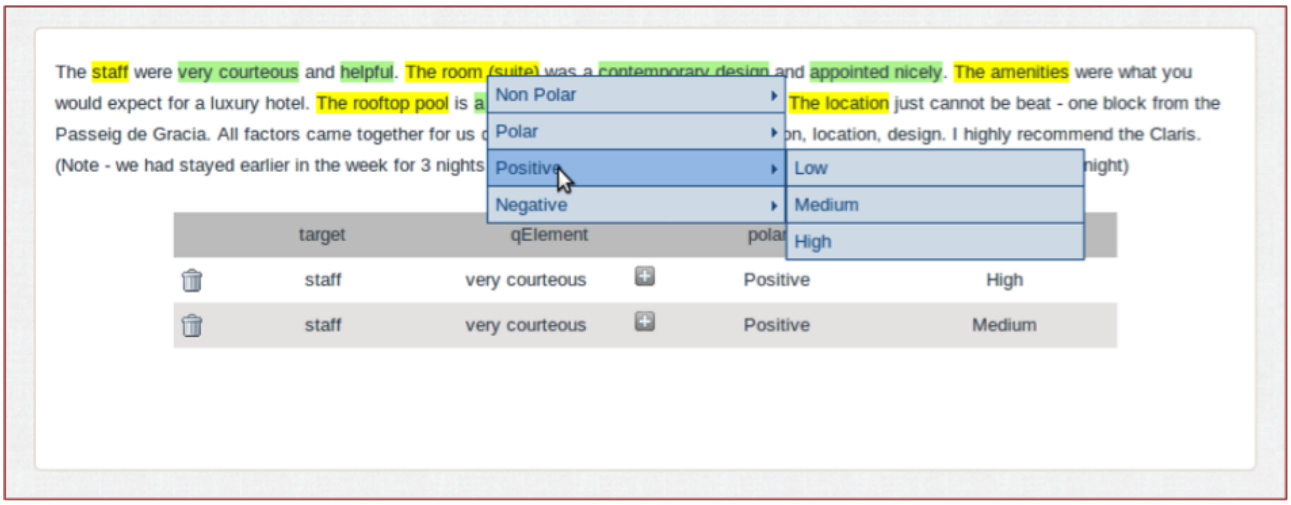
Annotation tools for relation-based tags: This type of tool appeared as necessary for classifying the polarity of opinion targets (in product reviews and tweets), since it appeared as basic to be able to associate to that target the cue elements supporting its polarity judgement.
The resulting annotation interface (running on CF) is shown in figure 1. In the screen, `targets` are shown in yellow and `cues` in green. The screenshoot captured the moment of assessing the `target` *"the room (suite)"* as positive. The next action would involve assessing the `intensity` of such polarity.

3.2. Strategies for on-the-fly controlling annotations quality

A crucial aspect in any annotation process is setting the mechanisms for dynamically controlling the quality of the resulting annotations. In annotation projects carried out by a small group of experts, such control is assured by means of activities like: training meetings at the beginning of the project for discussing partial results as well as the understanding of the annotation guidelines, collective annotation on the same text, individual but parallel annotation of the same text with a subsequent discussion on the challenges encountered, periodical meetings among the involved annotators, iterative refinement of the annotation guidelines based on the annotators feedback, etc.
In crowdsourcing-based annotation processes, however, it is not possible to implement these strategies of dynamic quality control, but both AMT and CF allow for other type of mechanisms.
CrowdFlower included a system for dynamically managing the annotations quality. Specifically, the customer (here, us) could include some of the instances to be tagged by the turkers already marked up with a gold annotation together with an explanation of why. These gold annotations were used to both train the turkers and to dynamically verify the correctness of their annotations. Based on that gold, during the annotation process the turker received immediate feedback on the answers he provided. If they differed substantially from the expected ones in the gold standard, his work was not accepted. At the same time, he received a message explaining why he was wrong, which helped him understand better the annotation guidelines.
However, this mechanism could only be applied on the standard templates provided by CF for for data classification. We were therefore able to use them only in the classification tasks, but not when using our own annotation tools for annotating text extents or setting relations between tags. That was one of the main drawbacks of using crowdsourcing-based platforms for opinion annotation. See section 5. for further considerations.
Amazon Mechanical Turk, on the other hand, did not counted on any system for either dynamically training the annotators or controlling of the quality of the resulting data, but it allowed the user to evaluate the gathered results before paying the turkers. The user could then accept the results (and pay the turker) or reject them (and not pay him). The evaluation on the gathered data could be undertaken in different ways: manually, against a gold standard, etc.
We created a gold standard for the tasks of classifying the tweets function (the only task we carried out through AMT). The gold standard represented 10% of the corpus, and contained only tweets with a clear-cut function. The annotations of a turker were accepted if she showed a 50% agreement with the gold standard. In addition, we established mechanisms to filter out fraudulent annotations. We observed that there were some annotators choosing almost systematically the same tweet function, or that spent an average time smaller than 2.5 seconds for classifying their tweets. We rejected those annotations, which were fed back to the annotation tool for other turkers to complete them.

3.3. Aggregation process

Each annotation layer was marked up by at least 7 turkers and only in some cases the agreement among them was complete. If the goal of the exercise were obtaining a distribution of the data (for instance, in order to reflect tendencies over the opinions), the set of (agreeing or disagreeing) values over each annotated token would be a perfectly valid result in itself (de Marneffe et al., 2012). However, the nature of the NewSoMe corpus had an applied side, with customers interested in receiving final values. Therefore, aggregating the multiple results over each token appeared as necessary.
CF had already a mechanism for aggregating data based on the gold standard supplied by the user and the confidence



The staff were very courteous and helpful. The room (suite) was a contemporary design and appointed nicely. The amenities were what you
would expect for a luxury hotel. The rooftop pool is a
The location just cannot be beat - one block from the
Passeig de Gracia. All factors came together for us
on, location, design. I highly recommend the Claris.
(Note - we had stayed earlier in the week for 3 nights
night)

	target	qElement		polar	
	staff	very courteous		Positive	High
	staff	very courteous		Positive	Medium

Figure 1: Annotation tool for relation-based tags

score of each turker obtained from comparing their performance against that. However, it only work on tasks carried out by means of their basic classification templates, which was not only the case in the creation of the NewSoMe corpora. On the other hand, AMT did not count (at least at that moment) with any aggregation method.

Aggregating classification tags: In CF-supported annotations, we took advantage of the aggregation mechanism facilitated by the system. In AMT-supported annotations (concerning only the tag `function` in tweets), we applied the strategy of the majority vote. In case of even results, we weighted it by means of the confidence score assigned by the system to each annotator.

Aggregating text-consuming tags: We interpreted this type of annotation as a binary classification task (token marked-up vs. not marked-up). Any token annotated by 2 or more annotators would be selected as marked-up in the final aggregation. A confidence mark would be in addition assigned to each token indicating the number of annotators that had tagged it: *plurality* (for agreement among 2 or 3 annotators), *majority* (4 to 5 annotators) and *absolute* (6 to 7 annotators).

Aggregating relation-based tags: As with the previous types of tags, we interpreted these annotations as classifications over pairs of targets and cues.

4. Data evaluation

Table 4 summarizes inter-annotation results based on the kappa κ metrics (Fleiss for multiple annotators) for the different annotation layers, sorted according to their annotation type: text-consuming (TEXT), classification (CLASS), and relation-based tags (REL). For each tag, it provides the κ average for all subcorpora in NewSoMe (e.g., blogs in English, tweets in Spanish, etc.), together with the minimum and maximum values.

The averages range between 0.18 and 0.51, with a minimum at 0.05 (for the tag `topic`, of type TEXT) and a maximum of 0.77 (tag `intensity` at the document level, of type CLASS). These are not excellent results, but are comparable to other work carried out by means of crowdsourcing platforms (Callison-Burch & Dredze, 2010).

Type	tag	Average	Min	Max
CLASS	`polarity` (doc)	0.50	0.41	0.58
	`intensity` (doc)	0.41	0.19	**0.77**
	`subjectiv.` (doc)	0.44	0.42	0.45
	`subjectiv.` (subdoc)	0.43	**0.18**	0.61
	`function`	0.28	0.21	0.35
TEXT	`topic`	0.31	**0.05**	0.60
	`segment`	0.18	0.08	0.28
	`target`	0.36	0.09	0.49
	`cue`	0.39	0.14	**0.75**
REL	`polarity` (subdoc)	0.51	0.21	0.72
	`intensity` (subdoc)	0.45	**0.08**	**0.73**

Table 4: Interannoation agreement results (in terms of κ).

In general, the results for TEXT tasks scored much lower than those for CLASS and REL tasks, which were conceptually easier since they provided the annotator with a list of classes where to choose from. Several of the highest scores reported in the Max column are from the expert annotations.

5. Lessons learned

Constraints on the degree of linguistic knowledge assumed. Certain layers of annotation appeared as particularly difficult for non-expert annotators. We experimented, for example, with annotating opinion holders (the person bearing an opinion) and the linguistic elements attributing the opinion to the holder (e.g., predicates such as *say, claim, argue*, etc.). These are common tags in opinion annotations of news reports (Kim & Hovy, 2006), and previous annotation projects on the same genre proved that they are easily identified by expert annotators (Wiebe et al., 2005; Saur & Pustejovsky, 2009). However, regardless of how devoid of linguistic terms the annotation guidelines were, the resulting data was always of very doubtful quality and so we decided to exclude it. The same happened with other elements such as intensifiers (particles of negation, modality, quantifiers, etc.), which have nevertheless been annotated as crucial in other annotation projects on subjective information (e.g., Vincze et al., 2008). Thus, the choice of the annotation method imposed constraints on the annotation scheme.

Not all languages were created equal. Whereas the annotations for the English and Spanish data were carried out within reasonable periods of time, the annotations on the Portuguese and Catalan sets were extremely slow, if not stagnant. Despite big efforts through social media channels for encouraging Catalan and Portuguese speakers to participate in our project, we had to hire two experts for each language to complete the process. In these cases, the effort invested in promoting annotation through crowdsourcing platforms and the time spent waiting for results ended up being more costly than the total amount spent for the final expert annotations.

The exponential effect. Previous work on resorting to crowdsourcing platforms for corpus annotation suggested that these can be adequate an adequate means for overcoming the issue of time (and, subsequently, money) cost of corpus development (e.g., Callison-Burch & Dredze, 2010; see however Fort et al., 2011). Nevertheless, results showed that the efficiency in annotating one single opinion layer on a monolingual corpus of a specific genre (e.g., Mellebeek et al., 2010) does not translate in linear terms to the annotation of multiple layers of information (target, cue, polarity, intensity, etc.) across several genres and languages, with its subsequent consequences in terms of the overall budget of the project.

Platform technical funcionalities Crowdsourcing platforms offer indisputable advantages to a great range of projects involving gathering massive amounts of data. They offer practical functionalities for obtaining data, computing the turkers reliability, applying on-the-fly mechanisms for quality control of the data, and aggregating multiple answers on the same item. These functionalities are nevertheless constrained to tasks of a certain nature, mainly item classification (regardless of their type: text, image, audio). Any other task differing from this type will not be able to take advantage of the whole set of functionalities otherwise available in the platform. This was a remarkable drawback in the project and required us to invest some effort in developing tools for carrying out a dynamic control of the data and the final aggregation.

Platform challenges. The crowdsourcing platforms posed other types of challenges that affected our project in a significant way:

AMT working policy: Customers of AMT services had to have a bank account in the USA. Initially, we arranged to use these services but later on we had to discontinue it because of administrative reasons and so had to move to CF.

Technical problems and bugs experienced with the CF platform. They involved a number of issues, from wrongly applying the gold standard, to incorrectly downloading the full set of results after task completion. In spite of the remarkable dedication and eagerness of the CF technical assistance to fix these problems, they ended up having a considerable impact in the overall duration of the project.

6. Related Work

Currently, most available opinion corpora are on English data, with only a few exceptions for example on Chinese (Ku et al., 2005, 2006), German (Li et al., 2012) and bilingual German and Spanish (Schulz et al., 2010). Furthermore, existing opinion corpora tend to focus on a specific genre or domain. In particular, news reports (Yu & Hatzivassiloglou, 2003; Wiebe & Riloff, 2005; Ku et al., 2005; Stoyanov & Cardie, 2008; Li et al., 2012), blogs (Ku et al., 2006), product reviews (Dave et al., 2003; Hu & Liu, 2004; Zhuang et al., 2006; Schulz et al., 2010; Toprak et al., 2010) or microblogs (Pak & Paroubek, 2010). Such narrow scope makes them perfect resources for developing mining tools and classifiers for the targeted genres, at the cost, however, of not being further applicable to other types of text or domains.

Some experiments on automatic corpus creation have proven quite successful (Wiebe & Riloff, 2005; Sarmento et al., 2009), but this area of work tends to be constrained to specific opinion dimensions (e.g., subjective vs. objective content or polarity detection). Recently, a new approach to corpus creation is being explored which involve using crowdsourcing platforms (Callison-Burch & Dredze, 2010). Our work fits within this paradigm.

7. Final remarks

The NewSoMe corpus provides additional annotated data to the area of opinion mining. It is significant due to its broad scope in terms of languages and genres covered, as well as the multi-layer annotation approach. Morevover, since all the genres covered there have been annotated under a unifying scheme, this resource contributes data through which to attain a more comprehensive understanding on how opinion is expressed in text.

Due to the wide-encompassing scope in its design, the NewSoMe corpus was developed by means of crowdsourcing platforms in order to mitigate time and money costs. Overall, however, we estimate that the whole project would have been at the most as costly (or possibly even cheaper) if run with expert annotators instead of resourcing to crowdsourcing platforms. The effort and time invested in overcoming the limitations, challenges and technical constraints of the platforms employed, and practical issues such as the scarcity of annotators for certain languages, were not compensated by the quality of the data obtained, even though it is comparable to that reported in similar corpus projects.

The NewSoMe set of corpora will be publicly available through the Linguistic Data Consortium.

8. Acknowledgments

The authors are very grateful to all the people that contributed to the creation of the NewSoMe corpus. Annotators: Estela Bayarri, Neise Mara Cardozo, Adriana Lopes, David Navarro. Annotation testers: Teresa Suñol, Yesika Laplaza, Montserrat Marquina, Guillem Massó. Technical assistants: David García Bautista, Bernat Grau. Manager: David Comas. We want to also thank the input received at different moments of the project by Carlos Rodríguez-Penagos, Joan Codina-Filbà, Jens Grivolla, Maite Melero, Marta Ruiz Costa-Jussà. The project was funded by means of a CENIT grant (CEN-20102037) by the Center for Industrial Technological Development (CDTI), Spanish Ministry of Economy and Competitivity, and a Marie Curie Reintegration Grant (PIRG04-GA-2008-239414) by the European Commission. Furthermore, it was sponsored by the companies Acceso, Neo Metrics and Yahoo! Iberia.

References

Callison-Burch, C. & Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, (pp. 112).

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, (pp. 519528).

de Marneffe, M.-C., Manning, C. D., & Potts, C. (2012). Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*, *38*(2), 301333.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413420.

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 168177).

Kim, S.-M. & Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text:* 18.

Ku, L.-W., Liang, Y.-T., & Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, (pp. 100107).

Ku, L.-W., Wu, T.-H., Lee, L.-Y., & Chen, H.-H. (2005). Construction of an evaluation corpus for opinion extraction. *Proc. of the Fifth NTCIR Wksp. on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, 513520.

Li, H., Cheng, X., Adson, K., Kirshboim, T., & Xu, F. (2012). Annotating opinions in german political news. In *LREC*, (pp. 11831188).

Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M. R., & Banchs, R. (2010). Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In *NAACL Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*.

Pak, A. & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC* 2010.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1135.

Sarmento, L., Carvalho, P., Silva, M. J., & de Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, (pp. 2936).

Saurí, R. & Pustejovsky, J. (2009). FactBank. a corpus annotated with event factuality. *Language Resources and Evaluation*, *43*, 227–268.

Schulz, J. M., Womser-hacker, C., & Mandl, T. (2010). Multilingual corpus development for opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta.

Stoyanov, V. & Cardie, C. (2008). Annotating topics of opinions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC'08*, Marrakech.

Toprak, C., Jakob, N., & Gurevych, I. (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*, (pp. 575584).

Vincze, V., Szarvas, G., Farkas, R., Mra, G., & Csirik, J. (2008). The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9(Suppl 11):S9*.

Wiebe, J. & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing* (pp. 486497). Springer.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, *39*(2-3), 165210.

Yu, H. & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, (pp. 129136).

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, (pp. 4350).