Annotating Expressions of Opinion and Emotion in the Italian Content Annotation Bank

Andrea Esuli[†], Fabrizio Sebastiani[†] and Ilaria C. Urciuoli[‡]

[†]Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via Giuseppe Moruzzi 1, 56124 Pisa, Italy
E-mail: ⟨*firstname.lastname*⟩@isti.cnr.it
[‡]Center for the Evaluation of Language and Communication Technologies (CELCT)
Via alla Cascata 56/c, 38050 Povo, Italy
E-mail: urciuoli@celct.it

Abstract

In this paper we describe the result of manually annotating I-CAB, the Italian Content Annotation Bank, by *expressions of private state* (EPSs), i.e., expressions that denote the presence of opinions, emotions, and other cognitive states. The aim of this effort was the generation of a standard resource for supporting the development of opinion extraction algorithms for Italian, and of a benchmark for testing such algorithms. To this end we have employed a previously existing annotation language (here dubbed WWC, from the initials of its proponents). We here describe the results of this annotation effort, including the results of a thorough inter-annotator agreement test. We conclude by discussing how WWC can be adapted to the specificities of a Romance language such as Italian.

1. Introduction

The Italian Content Annotation Bank (I-CAB) (Magnini et al., 2006) is a corpus of newspaper articles in the Italian language, manually annotated with semantic information of various types, including TEMPORAL EXPRESSIONS, different types of entities (such as PERSON ENTITIES, ORGANIZATION ENTITIES, LOCATIONS, and GEO-POLITICAL ENTITIES), and RELATIONS between such entities (such as, e.g., "affiliation", relating a person to the organization he/she is affiliated to)[1].

I-CAB was developed with the aim of making it both (a) a standard resource for supporting the development of algorithms for the automatic extraction of different types of information, and (b) a benchmark for testing such algorithms. Indeed, I-CAB has served as the reference resource and benchmark within several tracks of EVALITA'07, the recent campaign for the evaluation of NLP tools for the Italian language (Cappelli and Magnini, 2007).

In this paper we present our work on endowing I-CAB with a further level of (manual) annotation, i.e., *expressions of private state* (EPSs). A private state is "an internal state that cannot be directly observed by others", and as such includes "opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments" (Wiebe et al., 2005, pp. 168). Since opinion and emotion are arguably the two most important dimensions of private states, we will sometimes call (consistently with (Wiebe et al., 2005)) EPSs *expressions of opinion and emotion*. As such, our work squarely falls within the domains of *non-topical text analysys* and, more specifically, of *sentiment analysis*. This latter has received a lot of attention in the recent computational linguistics literature (see e.g., (Gamon and Aue, 2006)), also due to the increased applicative interest in the analysis of opinion- and emotion-laden language, as used in, e.g., political commentary, product reviews, and blogs.

A more recent addition to the set of subtasks of sentiment analysis is *opinion extraction*, the task of detecting, *within* a sentence or document, the expressions denoting the statement of an opinion, and detecting therein the subexpressions denoting the key components and properties (e.g., the opinion holder, the object of the opinion, the type of opinion, the strength of this opinion, etc.) within this statement (Breck et al., 2007; Choi et al., 2005; Choi et al., 2006; Kim and Hovy, 2005; Kim and Hovy, 2006). It is specifically this latter task, which lies at the crossroads of sentiment analysis and *information extraction*, which this work aims at contributing to, by providing a counterpart, in a language other than English, to a widely used English-language dataset for opinion extraction such as MPQA[2]. The very fact that I-CAB, aside from the annotations by EPSs, is also annotated according to the other dimensions cited above will not only stimulate research in opinion extraction from Italian texts, but will also provide a means for researchers to exploit potential synergies between different types of annotation.

This paper is structured as follows. In Section 2. we describe some relevant characteristics of I-CAB. In Section 3. we briefly sketch the annotation language for EPSs (hereafter dubbed WWC, from the initials of its proponents) that we have adopted from (Wiebe et al., 2005), describe the result of annotating the texts in I-CAB by means of it, and discuss potential areas of improvement for WWC as resulting from our experience.

2. I-CAB

I-CAB consists of 525 articles published by L'Adige[3], an Italian local newspaper, on four different days (September

[‡]Corresponding author. This work was done while the third author was affiliated with the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy.

[1] http://tcc.itc.it/projects/ontotext/icab.html

[2] http://www.cs.pitt.edu/mpqa/databaserelease

[3] http://www.ladige.it/

7, September 8, October 7, October 8, all in 2004). The articles are from the Current Events (87 articles), Cultural News (72), Economic News (54), Sports News (123), and Local News (189) sections of the (print edition of the) newspaper, and are subdivided into a training set of 335 articles (with an average length of about 339 word tokens) and a test set of 190 articles (with an average length of about 363 word tokens). For more details on I-CAB please refer to (Magnini et al., 2006).

3. Annotating Expressions of Opinion and Emotion in I-CAB

3.1. The WWC markup language

For annotating I-CAB by private states we decided to adopt an already available markup language, in order (a) to avoid "reinventing the wheel", (b) to leverage on past experience from other researchers, and (c) to ease comparisons between the same linguistic phenomena as occurring in different languages. This is completely in keeping with the policy adopted in (Magnini et al., 2006) for annotating I-CAB along the other dimensions discussed in Section 1., given that (Magnini et al., 2006) adopted markup languages previously developed within the ACE program[4].

We chose to adopt (what we here call) the WWC markup language developed in (Wiebe et al., 2005), since it was the result of the arguably most important annotation effort (the one which led to the development of the MPQA corpus) in the opinion extraction literature. In this section we present a brief introduction to WWC, referring the reader to (Wiebe et al., 2005) for a more detailed description.

WWC provides five types of *tags* (here indicated in SMALL CAPS). Each such tag can be further qualified by means of a number of *attributes* (here indicated in `typewriter` font). Aside from specifying in more detail the role played by the real-world entities denoted by the tagged expressions, attributes also allows to establish relations among the entities that play different roles in the same private state.

In WWC every EPS is mapped into a *private state frame*, i.e., a structured object in which the real-world entities that play a role in the EPS are annotated by means of the tags and further qualified by means of the attributes (see Table 1 for an example). In each private state a *source agent* holds a private state, optionally toward a *target agent*. WWC identifies three kinds of private states:

1. the explicit mention of a private state (e.g., "*I **fear** the Greeks, even when they bring presents*");
2. a speech event expressing a private state (e.g., "*You **said** you love her.*");
3. an expressive subjective element (e.g., "*He is a **nice** person*").

WWC also allows annotating *nested* EPSs in which the target agent is itself a private state (e.g.: "*John **wrote** me that Mary **said** I love pizza*"); the structured nature of private state frames naturally allows expressions at arbitrary levels of nesting to be represented.

A textual expression (*text span*, in WWC terminology) identifying the source agent or the target agent of a private state is annotated with the AGENT tag, which assigns a unique (at the document level) identifier to the entity denoted by the expression. Since EPSs can be nested, it is natural to identify the outermost source of every EPS in a given text as the author of the text itself; by convention, the identifier denoting the author of the text is "writer".

The explicit mention of a private state (Type 1 above), or a speech event expressing a private state (Type 2 above) are annotated using the DIRECT-SUBJECTIVE tag. The text span expressing either the mention of the private state or the speech event is identified, and the following attributes are specified:

- `intensity`: the intensity of the private state (`low` to `extreme`);
- `expressionintensity`: the contribution of the speech event expression to the intensity of the private state, e.g. "*say*" vs. "*cry*" (`neutral` to `extreme`);
- `insubstantial`: a Boolean flag indicating whether the private state is substantial to the discourse or not (e.g., hypothetical clauses are not substantial);
- `polarity`: the attitude of the private state, ranging on the values `positive`, `negative`, `other` and `none`;
- `nested` − `source`: the *chain* of agents expressing the private state;
- `target`: (optional) the agent which is the target of the private state.

The use of chains of agents to identify targets is the key WWC device for the expression of nested private state frames. For example, in the sentence "*John wrote me that Mary said I love pizza*" the DIRECT-SUBJECTIVE annotation related to the verb "*said*" has the target attribute equal to "writer/john/mary", because it is the author of the text who reports that John wrote that Mary said something about a private state.

Reported speech about objective facts is also annotated (e.g.: "*John **said** he is 30*"), using the OBJECTIVE-SPEECH-EVENT tag. A source agent and a target agent are assigned to the annotated text. WWC also includes an INSIDE tag, used for identifying the scope of a speech event (e.g.: "*Mary said **I love pizza***"). While this tag has not been used in MPQA (except for automatically marking an entire sentence as an INSIDE for "writer"), we have indeed used it in the annotation of I-CAB.

Finally, subjective expressions in text are annotated using the EXPRESSIVE-SUBJECTIVITY tag, that identifies the text span of a subjective expression and qualifies it by means of three attributes: source agents chain, intensity, and polarity of the expression.

3.2. Appying WWC to I-CAB

For annotating I-CAB by EPSs we have used the (freely available) GATE[5] tool developed at the University of

[4] `http://www.nist.gov/speech/tests/ace/`

[5] `http://gate.ac.uk/`

AGENT (text: “John”;
id: john;
nested − source: writer/john);
AGENT (text: “Mary”;
id: mary);
nested − source: writer/john/mary);
AGENT (text: “I”;
id: andrea;
nested − source: writer/john/mary/andrea);
AGENT (text: “pizza”;
id: pizza;
nested − source:
writer/john/mary/andrea/pizza);
DIRECT-SUBJECTIVE (text: “wrote”;
intensity: low
expressionintensity: neutral;
polarity: positive;
insubstantial: false;
nested − source: writer/john;
target: mary);
INSIDE (text: “Mary said I love pizza”;
nested − source: writer/john);
DIRECT-SUBJECTIVE (text: “said”;
intensity: low;
expressionintensity: neutral;
polarity: positive;
insubstantial: false;
nested − source: writer/john/mary;
target: andrea);
INSIDE (text: “I love pizza”;
nested − source: write/john/mary);
DIRECT-SUBJECTIVE (text: “love”;
intensity: high;
expressionintensity: high;
polarity: positive;
insubstantial: false;
nested − source: writer/john/mary/andrea;
target: pizza)

Table 1: The private state frame generated by the sentence “*John wrote me that Mary said I love pizza*”.

Sheffield (Cunningham, 2002). This is unlike (Magnini et al., 2006), who for the other annotations of I-CAB had instead used the Callisto system, developed by MITRE Corporation. The reason we have chosen GATE is that, since this was the system originally used in (Wiebe et al., 2005), WWC was already available on it, thus sparing us of the additional effort required in customizing an annotation tool to it.

Consistently with the other types of annotation on I-CAB described in (Magnini et al., 2006), our EPSs annotations are encoded in MEAF (Bentivogli et al., 2003), an XML-based format compliant with the guidelines set by the Text Encoding Initiative (TEI). However, since the annotations generated by GATE are not in MEAF, we had to implement a translator from the format generated by GATE into MEAF.

One of the advantages of having all types of annotations on I-CAB expressed in the same format is that it allows us to interlink them and, navigating across the various types, to discover new relevant information. For example, connecting AGENT annotations with named entities annotations, and the using the coreference information on named entities, enables us to find all the EPSs in which a given named entity plays some role.

Some quantitative results of the annotation process are reported in the first column of Table 2, in which the number of annotations for each of the five tags is reported.

3.3. Inter-Annotator Agreement

In order to test whether (a) the annotation produced is high-quality and (b) whether the meaning of the tags in WWC is uncontroversial, we conducted an inter-annotator agreement (IAA) study. We asked an intern (a third-year student in Computers and the Humanities) to independently annotate 127 (94 training and 33 test) articles of I-CAB (this accounts for 24% of the total 525 articles). Prior to doing this, we annotated 10 (7 training and 3 test) articles together with the intern, so as to align his and our interpretations of the various tags.

Concerning how IAA should be measured note that, for each individual tag, annotation can be viewed as an instance of binary classification (since annotation amounts to deciding whether a given expression is or is not an instance of the tag); this means that measures for binary classification accuracy can be used to measure IAA, and vice versa. Two examples of such measures are F_1 (Lewis, 1995), from the tradition of binary classification, and Cohen's κ (Cohen, 1960; Di Eugenio and Glass, 2004), from the tradition of IAA. Each of them consists in a function computed over the confusion table T, i.e., a 2×2 table in which each of the four entries lists the number of objects that coder A / coder B have deemed to be instances / non-instances of the tag, in each of the four possible combinations.

However, in order to measure IAA we should also specify who the objects of classification are. (Wiebe et al., 2005) define the *overlap* measure of IAA, where the objects of classification are all the expressions which either A or B have annotated with the tag, and overlap is defined as the average of the agreement between coder A and coder B and the agreement between coder B and coder A on such objects (where their *agreement* measure is a non-symmetric measure that counts how many of the expressions annotated by one coder have also been annotated, at least partially, by the other coder).

However, following (Esuli et al., 2008), we think that this IAA measure is too coarse, since it gives partial credit to partial agreement (defined as the case in which the two annotators annotate overlapping but non-coinciding portions of text by means of the same tag) without taking into account the *degree* of this overlap. For instance, two annotations that are each 10 words long and that overlap by 1 word only receive the same partial credit as two annotations that are each 10 words long and that overlap by 9 words, which is unintuitive. (Esuli et al., 2008) have thus devised what they call *the Token Model* of IAA (in contrast to (Wiebe et al., 2005)'s “annotated expressions” model, or AnnExp as

we call it in Table 2), in which each token (i.e., word) in the text is viewed as the object of a classification decision (i.e., the decision whether the token falls within the text span annotated by the tag). Having individual tokens, rather than possibly complex expressions, as the objects of classification, allows one to implement a much more fine-grained view of IAA.

(Esuli et al., 2008) have subsequently introduced a further refinement of this model, called the *the Token&Blank Model* of IAA, which is justified by a small shortcoming of the token model, i.e., by the fact that the Token Model would not distinguish between annotating a portion of text consisting, say, of two consecutive words, as a single annotation of length 2 or as two distinct annotations of length 1. As a consequence, in the Token&Blank Model the objects of classification are all the words in the text *and* all the blanks (actually: all separators) in the text. In such a way, annotating a portion of text consisting of two consecutive words as a single annotation of length 2 will entail annotating both words *and their separating blank too*, while annotating it as two distinct annotations of length 1 will entail annotating both words *but not their separating blank*. Detailed mathematical definitions of the three IAA models used here are given in (Esuli et al., 2008).

The results of our IAA study are reported in Table 2; Annotator A is the third author of this paper, while Annotator B is our intern.

It is immediate to note that the numerical results of the Token Model and of the Token&Blank Model are very close to each other, for both the κ and F_1 measures. This could be expected, since a substantial numerical difference between the results of the two models could only be the result of systematic inter-annotator disagreement on whether the same sequence of words must consist of a single long annotation or of multiple shorter annotations with the same tag, which seems unlikely at best.

A second observation that can be made is that some tags are much more controversial than others, as witnessed by very different levels of IAA. For instance, there seems to be pretty high agreement between the annotators on the INSIDE tag, while the agreement on the EXPRESSIVE SUBJECTIVITY tag is generally much lower. Agreement on the other three tags (AGENT, DIRECT SUBJECTIVE, and OBJECTIVE SPEECH EVENT) is somehow intermediate between these two extremes. That EXPRESSIVE SUBJECTIVITY is controversial can also be seen by the sheer number of annotated sequences, a figure that for annotator A is almost double as for annotator B (924 vs. 467); in other words, A frequently sees subjectivity where B does not see it. This is not surprising, since the very notion of subjectivity is elusive: does the sentence "*It will take some time before things improve*" have a subjective character? Annotator A thought it contained a grim, pessimistic statement, while annotator B thought it depicted a fairly neutral assessment.

A third observation is that the AnnExp Model can differ significantly from the Token Model and Token&Blank Model in its results, as witnessed for example by the EXPRESSIVE SUBJECTIVITY tag, in which the first yields good results while the second and third ones produce very low figures. This seems to indicate that, when two annotations, from two different annotators, with the EXPRESSIVE SUBJECTIVITY tag overlap, the degree of this overlap tends to be very low, since the two models sensitive to this degree produce a low score.

3.4. How adequate is WWC for representing EPSs?

The WWC markup language, together with the guidelines for using it, presented in (Wiebe et al., 2005) was originally developed with the English language in mind. One of the aims of our work was also to test to what extent WWC proved adequate to dealing with other languages (and, specifically, Italian, which is morphologically much richer than English) with possibly different morpho-syntactic characteristics.

One problem which WWC proved inadequate in solving is the fact that Italian, unlike English, allows the direct or indirect object personal pronoun to appear as a *clitic*, i.e., an element whose grammatical status is somewhere in-between a typical word and a typical affix (e.g., "*dammi*" (= "*give me*"), or "*dammelo*" (= "*give it to me*")). The former example is actually an instance of an *enclitic* pronoun, since the pronoun "*-mi*" is a suffix to the stem of the host word "*dam-*", while in the latter example the pronoun "*-me-*" is a *mesoclitic*, since it appears between the stem of the host word and another affix. This feature characterizes Romance languages in general. The main consequence is that, for Italian and other Romance languages, annotation would be best carried out at the morphosyntactic level rather than at the orthographic level. In I-CAB we do not annotate clitic pronouns, and we instead link all the annotations referring to them to the closest non-clitic mention of the referred entity.

A second problem that needs to be taken into account is the fact that Italian allows the subject of a sentence to be implicit: e.g., "*Io ho mangiato*" (= "*I have eaten*") and "*Ho mangiato*" (same) are equivalent, equally acceptable sentences, the subject of the first being only implicit. In the common case in which the subject is mentioned in one of the preceding (or succeeding) sentences, we solve this problem by linking all the annotations referring to the implicit subject to its closest mention. In the rare case in which the implicit subject is never mentioned in the entire document, we link all the annotations referring to it to a `writer/X` agent id, where `X` is a unique id never used in the rest of the document.

A third problem we detected is that text spans that play a given role in an EPS must consist, according to WWC, of a single contiguous piece of text. However, this is not always the case in real text. For instance, in the sentence "*It will take some time – the colonel said – before things improve*", the non-contiguous portion of text "*It will take some time (...) before things improve*" should play the INSIDE role in the scope of "*the colonel said*". In order to solve this problem we annotate the two fragments as two distinct INSIDE text spans, instead of including irrelevant text into the span. We then add two optional attributes to the INSIDE tag, `id` and `link`, in order to express the interdependency between the two annotations:

INSIDE (`text` :"It will take some time";
`id` : `in1`; `link` : `in2`;

	# of annotations		AnnExp		Token		Token&Blank	
	A	B	AGR	F_1	κ	F_1	κ	F_1
AGENT	1239	859	.539	.521	.442	.481	.439	.472
DIRECT SUBJECTIVE	263	246	.507	.507	.432	.442	.414	.422
EXPRESSIVE SUBJECTIVITY	924	467	.602	.537	.370	.392	.339	.357
INSIDE	491	563	.767	.763	.717	.793	.718	.791
OBJECTIVE SPEECH EVENT	132	144	.501	.500	.471	.476	.462	.465

Table 2: Number of annotations for the various tags (first 2 columns), and results of the IAA study according to various IAA models (remaining columns).

`nested − source`: writer/colonel;)
INSIDE (`text` :"before things improve";
`id` : `in2`; `link` : `in1`;
`nested − source`: writer/colonel;)

We should note, however, that these reported problems should not be understood as a negative critique of WWC, since this was admittedly not meant to be a complete set of primitives for describing *all* the devices used in natural language for expressing PSs. The array of such devices is certainly bewildering, and, as stated in (Wiebe et al., 2005), "[the WWC] annotation scheme covers *a broad and useful subset of* the range of linguistic expressions and phenomena employed in naturally occurring text to express opinion and emotion." (our emphasis)

Acknowledgments

This work was partially supported by Project ONTOTEXT "From Text to Knowledge for the Semantic Web", funded by the Provincia Autonoma di Trento under the 2004–2006 "Fondo Unico per la Ricerca" funding scheme. We thank Riccardo Sagratini for participating in the inter-annotator agreement study, and the other members of Project ONTOTEXT for useful discussions.

4. References

Luisa Bentivogli, Christian Girardi, and Emanuele Pianta. 2003. The MEANING Italian corpus. In *Proceedings of the Conference on Corpus Linguistics (CL'03)*, pages 103–112, Lancaster, UK.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, pages 2683–2688, Hyderabad, IN.

Amedeo Cappelli and Bernardo Magnini, editors. 2007. *Proceedings of the 1st Workshop on the Evaluation of NLP Tools for Italian (EVALITA'97)*. Roma, IT.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 355–362, Vancouver, CA.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 431–439, Sydney, AU.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Andrea Esuli, Michał Pryczek, and Fabrizio Sebastiani. 2008. New external measures for the evaluation of information extraction systems. Forthcoming.

Michael Gamon and Anthony Aue, editors. 2006. *Proceedings of the ACL'06 Workshop on Sentiment and Subjectivity in Text*. Sydney, AU.

Soo-Min Kim and Eduard Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of the AAAI'05 Workshop on Question Answering in Restricted Domains*, Pittsburgh, US.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the ACL'06 Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sidney, AU.

David D. Lewis. 1995. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95)*, pages 246–254, Seattle, US.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi-Lenzi, and Rachele Sprugnoli. 2006. I-CAB: The Italian content annotation bank. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 963–968, Genova, IT.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):165–210.