SentiML: functional annotation for multilingual sentiment analysis

Marilena Di Bari
Centre for Translation Studies
University of Leeds
Leeds, UK
mlmdb@leeds.ac.uk

Serge Sharoff
Centre for Translation Studies
University of Leeds
Leeds, UK
s.sharoff@leeds.ac.uk

Martin Thomas
Centre for Translation Studies
University of Leeds
Leeds, UK
m.thomas@leeds.ac.uk

ABSTRACT

Sentiment Analysis is the task of automatically identifying whether a text or a single sentence is intended to carry a positive or negative connotation. The commonly used Bag-of-Words approach that relies on counting positive and negative words, whose connotation is indicated by specially crafted sentiment dictionaries, is not ideal because it does not take into account the relations between words and how the connotation of single words changes according to the context. This paper proposes a way of identifying and analysing the targets of the opinions and their modifiers, along with their linkage (appraisal group) through an annotation schema called SentiML. Such schema has been developed in order to facilitate the identification of these elements and the annotation of their sentiment, along with advanced linguistic features such as their appraisal type according to the Appraisal Framework. The schema is XML-based and has been also designed to be language-independent. Preliminary results show that the schema allows more coverage than a sentiment dictionary, while achieving reasonably fast and reliable annotation in spite of its fine granularity.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—Linguistic processing

General Terms

Annotation schema

Keywords

Sentiment analysis, Appraisal Theory

1. INTRODUCTION

Recent years have seen the increase in the amount of data in electronic format available on the Internet in a variety of contexts, including opinions on products, services, individuals and issues [9]. As a consequence, a brand new field called *Sentiment analysis* has been rapidly growing with the aim of classifying such sentiments and attitudes in classes, e.g., positive and negative.

DH-case '13, September 10 2013, Florence, Italy

http://dx.doi.org/10.1145/2517978.2517994

However, opinions are often not expressed as simple and direct assertions in which counting the number of positive or negative words suffices. They contain a number of syntactic and stylistic devices that current automatic systems are not fully able to deal with, e.g., negation, polarity modification, sentiment words used in a non-sentiment context, or opinions only implied rather than explicitly stated.

For example, the election pledge *"I will eliminate capital gains taxes for the small businesses"* is likely to be expressed with a positive sentiment, but it would be incorrectly classified as negative because of two apparently negative words (*eliminate* and *taxes*). What actually happens is that the verb *eliminate* acts as polarity reversal for the noun *taxes*. In addition, quite often opinions on different aspects of a topic can be found in the same review or even the same sentence. For example, when giving an opinion on films, users express sentiments about the plot, the actors' performance or director's choices. Sometimes praising the plot is even correlated with a negative evaluation overall.

In our research, we tackle such problems by proposing a schema, called *SentiML*, which allows multi-level annotations of three categories: *target* (expression the sentiment refers to), *modifier* (expression conveying the sentiment) and *appraisal group* (couple of modifier and target). The sentence previously considered will thus be annotated in the following way:

> "I will $[[\text{eliminate}]_{\text{M}}$ capital gains $[\text{taxes}]_{\text{T}}]_{\text{AG}}$ for the small businesses"

where *T*, *M* and *AG* respectively refer to target, modifier and appraisal group. Once the group with its target and modifier has been identified, its orientation and appraisal type are marked by using specific attributes.

Sentiment annotation at the sub-sentence level has previously been reported in the literature [8, 18], showing significant improvements in the performance of tasks such as the extraction of opinion expressions along with their holders and polarities. However, these studies have generally been performed on a rather ad hoc basis, while the schema proposed here is based on the *Appraisal Framework* (AF) [10], an established linguistically-grounded theory with a range of applications.

1.1 The Appraisal Framework

The Appraisal Framework is a development of *Systemic Functional Linguistics* [6] concerned with the study of the language of evaluation, attitude and emotion in written texts.

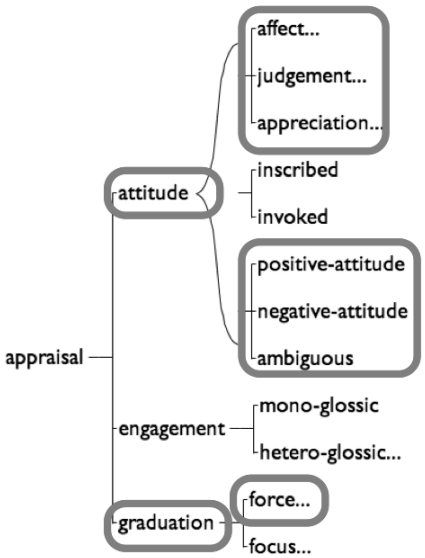

Figure 1: Excerpt of the Appraisal framework. Highlighted concepts are those taken into account in this research.

It consists of three sub-systems that operate in parallel: *attitude*, *engagement* and *graduation*. Of these, *attitude* is the most studied and is sub-divided into *affect*, which deals with personal emotions and opinions (e.g., *happy*, *sad*); *judgement*, which concerns author's attitude towards people's behaviour (e.g,. *heroic*, *craven*); *appreciation*, which considers the evaluation of things (e.g., *ugly*, *useful*). The *engagement* sub-system considers author's attitude towards the behaviour of other people, whereas the *graduation* sub-system investigates how the use of language amplifies or diminishes attitude and engagement. In particular, *force* is related to intensity, quantity and temporality. An excerpt of the general system is depicted in Figure 1.

The AF has been widely used in linguistics and, most recently, in the field of sentiment analysis, because it provides a robust way of identifying the targets of the opinions and their attributes. Even in the case in which the complexity of the frames has been simplified by using the appreciation value always for things, the judgement value always for persons and the affect value always for oneself [16], it has been pointed out that the same adjective can take a value rather than the other depending on the context in which it is used. For example, the adjective *good* expresses positive affect in "good feeling" , positive judgement in "good parent", and positive appreciation in "good book" [12]. Starting from the AF, the concept of *appraisal expression* has been also defined [2] as an expression including a *source*, an *attitude*, and a *target*, each represented by various attributes. For example, in *"I found the movie quite monotonous"*, the speaker ("I") expresses a negative attitude ("quite monotonous") towards the target ("the movie"). While early works have focused on appraisal expressions consisting of adjectives only [2, 17, 1], recent works have taken into account other grammatical categories as well [14, 12].

2. SENTIML SCHEMA

We unified the works mentioned in the previous section by proposing an annotation schema that is: (I) designed to allow fast multilayer annotation; (II) applicable to *appraisal groups* rather than more complex expressions; (III) flexible enough to consider groups consisting of grammatical categories other than adjectives (i.e. nouns, verbs, adverbs and pronouns); (IV) designed to be applied to different languages. The final aim is to study how appraisal is conveyed in the social, political and economic spheres within a given culture, as well as across different cultures.

The proposed schema consists of three categories: **targets, modifiers** and **appraisal groups**. They will be presented in the following sections.

2.1 Target

A target is any entity (object, person or concept) that is implicitly or explicitly regarded as positive or negative by the author of the text. The following scenarios can be found:

- One target with one feature. For example, in *"This article is useless"*, there is one target (*article*) with a feature (*useless*).
- Different targets with their features. For example, in *"I bought a good camera from that website, but a very bad phone"*, there are two targets (*camera* and *phone*) and their related features (*good* and *bad*).
- Different targets with their features referring to one entity. For example, in *"The camera has a good quality, but a considerable weight"*, the two targets (*quality* and *weight*) have got their own features (*good* and *considerable* respectively), but they refer to the same object (*camera*).

In terms of its logical function, a target can be either a *subject* or an *object*, depending on which carries the sentiment (e.g., in "we share beliefs", only *beliefs*). From a grammatical point of view, it can be a common noun (e.g., *children*, *cat*), a proper noun (e.g., *Olivia*, *China*, *Sparky*), a personal pronoun (e.g., *they*) or a verb (e.g., "*to talk* too fast").

Targets have two attributes:

Type. This attribute captures the type of target and has five possible values: 'person', 'thing', 'place', 'action' and 'other'. Animals are included in the category 'thing'. Countries, cities, provinces and natural geographical points (e.g., rivers, lakes, mountains) are usually annotated as 'place', whereas *world* can be either 'thing' or 'place', depending on whether it carries the action or not. The 'other' value is only used when an adjective is marked as target (e.g., "$[\text{easily}]_{\text{M}}$ $[\text{imaginable}]_{\text{T}}$").

Orientation. This attribute captures the *prior* orientation of a target and it has four possible values: 'positive', 'negative', 'neutral' and 'ambiguous'. For example, *peace* is positive regardless of the context, whereas *pessimism* is negative regardless of the context. The 'ambiguous' value is given if the orientation depends on the context (e.g., the word *challenge* in "promising challenge" and "unfair challenge"). In this case, the appropriate orientation is annotated in the appraisal group (e.g., "promising challenge" is marked as positive and "unfair challenge" is marked as negative). Quite common is also the case in which words do not seem ambiguous at first (e.g., *growth* can give the idea of being always positive, but it is actually negative when modified, for example, by *slow*). The 'neutral' value is assigned to targets that have no connotation, very often personal pronouns (e.g., *we*) and places (e.g., *America*).

2.2 Modifier

A modifier is what *modifies* the target. It can be an adjective (e.g., "$[\text{beautiful}]_{\text{M}}$ $[\text{car}]_{\text{T}}$"), a verb (e.g., "$[\text{obtain}]_{\text{M}}$ $[\text{victory}]_{\text{T}}$"), an ad-

verb (e.g., "[foolishly]$_M$ [sought]$_T$") or a noun (e.g., "[alliance]$_T$ for [progress]$_M$", which is equivalent to say "progress alliance").

Modifiers have four attributes:

Orientation. This attribute refers to the *prior* orientation of a modifier and it has four possible values: 'positive', 'negative', 'neutral' and 'ambiguous'. For example *beautiful* is positive regardless of the context, whereas *horrid* is negative regardless of the context. The 'ambiguous' value is used when the orientation depends on the context (e.g., for the verb *wishes* in the cases "wishes well" and "wishes ill"). Sometimes the same word takes a different orientation according to the sense (e.g., *light* used as modifier in "light the world" is positive, whereas *light* used as target in "give light" is ambiguous).

Attitude. According to the AF (see Section 1.1), attitude has three possible values: 'affect', 'judgement' and 'appreciation'. The 'affect' value is used for personal states (e.g., "I'm optimistic") and opinions (e.g., "if we are divided, we won't achieve much"). The 'judgement' value is used for others' behaviour (e.g., "children are unwilling to obey"), whereas the 'appreciation' value is used for the evaluation of things (e.g., "solemn oath").

Polarity. This attribute captures the information linked to the presence of a negation. It has two possible values: 'marked' and 'unmarked'. It is 'marked' when there is a negation (e.g., "we do not observe a victory"), 'unmarked' otherwise (e.g., "we like this place") . Apart from being local (e.g., "not good"), negation can also involve long-distance dependencies (e.g., "does not look very good") or the subject (e.g., "no one thinks that it is good") [18].

Force. This attribute refers to the intensity of the modifier. It has four possible values: 'high', 'low', 'normal' and 'reverse'. The 'high' value is used if the modifier is preceded by adverbs of high intensity such as *very* and *extremely* (e.g., "very good", "extremely good"), if it is included in expressions such as "not only ..., but ..." (e.g., "not only good, but amazing") or if it expresses high intensity itself (e.g., *best* as opposed to *good*). In Figure 2 the adverb *definitely* gives high force to the modifier. The value 'low' is used if the modifier is preceded by adverbs of low intensity such as *less*, *little* and *poorly* (e.g. "less good") or it expresses low intensity itself (e.g., *worse* as opposed to *bad*). The value 'normal' represents the standard input and is used when the modifier is not modified by anything (e.g. *good)*. The 'reverse' value is used in presence of words called *reversals* because they reverse the prior orientation of their targets (e.g., the verb *abolish* in "abolish taxes"). Apart from verbs (e.g., *to decrease*, *to limit*, *to diminish*, *to remove*), reversals can be nouns (e.g., *termination*), prepositions (e.g., *without*, *despite*) and condition operators (e.g., *if*, *even though*).

2.3 Appraisal group

An appraisal group represents an opinion on a specific target. For this reason, it is defined as the link between the target and the modifier. In SentiML it always matches one of the following combinations:

- A noun with an adjective. For example, "[good]$_M$ [plan]$_T$".
- A pronoun with a noun. For example, in a relational clause like "[they]$_T$ [cliches]$_M$".
- A noun with a noun when linked by prepositions. For example, "[stigmatization]$_T$ of [people]$_M$". The usual prepositions are *of*, *for*, *in*, *against*, *with*, *towards*, *between*.
- A verb with an adverb. For example, "[strongly]$_M$ [support]$_T$.
- A noun with a verb. For example, "[children]$_T$ [love]$_M$.

Appraisal groups have just one attribute:

Orientation. This attribute refers to the *contextual* orientation of the appraisal group and it has four possible values: 'positive', 'negative', 'neutral' and 'ambiguous'. For example, "[hungry]$_M$ [minds]$_T$" has a positive contextual orientation, whereas "[hungry]$_M$ [children]$_T$" a negative one if referred to the Third World. The 'ambiguous' value is assigned when the appraisal group does not have a clear orientation (e.g., "[increasing]$_M$ [demand]$_T$"). The 'neutral' value is actually never used during the annotation, as we are interested only in targets which carry sentiment (see Section 1 for further details). It is important to underline that the prior orientation of a word might change when embedded in a group, for example *lack* might seem negative, but is positive in "[lack]$_T$ of [pain]$_M$". Finally, contextual orientation can also be influenced by domain/topic (e.g., *cool* is positive for car, but negative for behaviour) [18].

3. CORPORA

We decided to apply SentiML to a wide range of argumentative texts:

- **Political speeches**. Mainly American presidents' addresses available on the web[1].
- **Talks**. *TED* (Technology, Entertainment, Design) talks[2] on a wide range of topics under the slogan "ideas worth spreading".
- **News**. Belonging to the "human rights" domain of the *MPQA opinion corpus* [18]. This corpus contains texts with manual annotations of opinions and private states such as beliefs, emotions, sentiments and speculations.

Many materials in these categories are available in languages other than English and would easily allow research on inter-cultural and cross-cultural critical discourse analysis. For example, some of the American presidential addresses have been translated into other languages, and they can be easily compared with the examples of original texts. Many of the TED talks have been translated into other languages and made available with their alignments in the WIT^3 corpus [3].

In terms of news, we decided to consider those from the MPQA corpus [18] although in English only, in order to allow a future comparison of the performance of the automatic annotation done according to SentiML to that done according to the MPQA. The major differences are that: (1) in the MPQA annotation there is no distinction between the AF appraisal types (i.e. affect, judgement and appreciation), but there is only one big category called *sentiment*; (2) each group (called *attitude frame*) covers a much larger span than ours; (3) implicit links are also marked, for example when quoted speech are not preceded by *said*; (4) a *both* tag is used when both a positive and negative sentiment are being expressed (e.g., "a bittersweet memory"). News originally produced in other languages or for which a translation is available will be considered as well.

[1] `http://avalon.law.yale.edu/subject_menus/inaug.asp`
[2] `http://www.ted.com/talks`

```
<?xml version="1.0" encoding="UTF-8" ?>
<AppraisalAnnotation>
  <TEXT><![CDATA[
    They definitely sparked outrage.
  ]]></TEXT>
  <TAGS>
    <APPRAISALGROUP id="A0" fromID="M0" fromText="sparked" toID="T0" toText="outrage"
                    orientation="negative" />
    <MODIFIER id="M0" start="21" end="28" text="sparked" attitude="judgement"
              orientation="ambiguous" force="high" polarity="unmarked" />
    <TARGET id="T0" start="29" end="36" text="outrage" type="thing"
            orientation="negative" />
  </TAGS>
</AppraisalAnnotation>
```

Figure 2: Example of the sentence "They definitely sparked outrage" annotated in SentiML format

Studies with topics similar to ours, such as the style of political leaders [5], their ideological positions [13], reactions to political debates and campaigns in political blogs [4] did not make use of the AF. Studies based on the AF were instead confined to reviews, and some did not study granularity of sentiment expressions. Taboada and Grieve [16] automatically extracted adjectives from 400 reviews (200 positive and 200 negative) on different topics (books, movies, music, phones, cars and cookware). However, their system was not so accurate when applied to product reviews where various components of the same object were evaluated. For example, in car-related reviews, components such as acceleration, safety and appearance were often mentioned. The problem with their system was that, even when some of these parts were evaluated as positive, the negative tag was assigned. In movie reviews they instead assumed that the overall evaluation was reflected in each one of the components (e.g., score, plot, director, actors) or that comments on each component were used to support a negative/positive classification. Whitelaw et al. [17] automatically gathered 1329 adjectival modifiers by starting from 400 seed terms, through Wordnet and two thesauri. However, this allowed them to cover only their dictionary orientation, not their contextual one. Read and Carroll [14] asked the annotators to annotate 1245 sentences with no precise guidelines of what the span should have been, thus increasing the complexity of comparison and reducing the inter-annotator agreement. Zhao et al. [19] annotated a similar number of appraisal expressions to ours (894), where by *expressions* they meant *groups*. Only the domains were different, in so far as they concentrated on cameras and MP3 player reviews. They built syntactic linkages between polarity words and their targets by using dependency parsing, and afterwards generalised such linkages according to their part-of-speech (e.g, "The camera's image is perfect" had the same path as "The camera's image would be perfect"). They demonstrated that this mining of syntactic relationships between polarity words and their targets was more helpful for appraisal recognition than the association of a word and a target based on their proximity.

4. ANNOTATION WORKFLOW

We designed SentiML as an XML-based schema and we used MAE [15], a freely available annotation environment, to implement the annotation. MAE keeps the annotations separate from the content of the input documents by using stand-off annotations, in accordance with the *ISO Linguistic Annotation Framework* (LAF) [7]. Figure 2 shows the XML output of the annotation of the expression *"They definitely sparked outrage"*. The annotation contains: (I) one appraisal group (*"sparked outrage"*) with orientation value 'negative', (II) one modifier (*sparked*) with attitude value 'judgement', orientation value 'ambiguous', force value 'high' and polarity value 'unmarked', (III) one target (*outrage*) with type value 'thing' and orientation value 'negative'.

The Document Type Definition (DTD) of SentiML to be used with MAE, which includes the categories and the attributes described in Section 2, is publicly available[3].

4.1 Special cases

In all the domains under analysis, we found that there were cases in which the adverb radically modified the force of the group (e.g., "those who foolishly sought power"). For the purpose of finding the best way to deal with them, we tried two different annotation styles:

- **Complete**. A new group was created ("$[\text{foolishly}]_M$ $[\text{sought}]_T$") apart from the main one ("$[\text{sought}]_M$ $[\text{power}]_T$"), since it fell into one of the combinations mentioned in Section 2. In the first one, *foolishly* had 'high' force and *sought* had 'normal' force.
- **Light**. The verb in the main group ("$[\text{sought}]_M$ $[\text{power}]_T$") took the force of the adverb *foolishly*.

We finally opted for the 'complete' style, since it required no significant additional effort in comparison with the 'light' one.

We also found a number of complex cases to annotate, which we will now list.

More than one modifier. If a target has more than one modifier, as in the expression "cultural and spiritual origins", one group for each modifier is created ("$[\text{cultural}]_M$ $[\text{origins}]_T$", "$[\text{spiritual}]_M$ $[\text{origins}]_T$").

Phrasal/multi-word verbs. These are verbs followed by a particle or a preposition forming a single semantic unit (e.g. "cast off"). They are annotated as single tokens and embedded in a group afterwards (e.g., "$[\text{cast off}]_M$ $[\text{worries}]_T$"). Other examples of phrasal verbs are "get rich", "put forward", "worry about". In case the phrasal verb is split (e.g., "*carried* us *up*", "*tore* the world *apart*"), in MAE it is possible to annotate either the verb (e.g., *carried*) or the preposition (e.g., *apart*) if they make sense on their own. Multi-word verbs are verbs that can use also a noun to convey a concept (e.g., "symbolizing an end", "signifying renewal", "proclaim an

[3] `http://corpus.leeds.ac.uk/marilena/SentiML/AppraisalAnnotation.dtd`

end").

Multi-word expressions. These are annotated as single tokens (e.g. "at issue", "at odds", "in practice", "out of control", "under pressure", "in the light", "go off a cliff"). They include the case of idioms and fixed expressions such as "on the cutting edge", which would be annotated as "[cutting]$_M$ [edge]$_T$" and "settling for less" annotated as "[settling]$_T$ [for less]$_M$". In the case of words not placed next to each other, like *never* and *before* in the expression "never seen our planet from this perspective before", they are not annotated.

Co-reference. When an element is referring to something else mentioned before or after, as in the short paragraph *"Let's begin with some images. They're iconic, perhaps cliches"*, the pronoun *they* rather than the actual subject *images* is annotated in two groups ("[they]$_T$ [iconic]$_M$" and "[they]$_T$ [cliches]$_M$").

Non-sentiment words. We do not annotate words that carry no sentiment in a given context, even where they feature in a sentiment dictionary. For example:

- "Worldly possessions" in *"They packed up their few worldly possessions and traveled across oceans in search of a new life"*.
- "Double standards" in *"No double standards should be pursued here"*.
- "Double standard" in *"America has a double standard policy"*.
- "Scale of our ambitions" in *"There are some who question the scale of our ambitions – who suggest that our system cannot tolerate too many big plans"*.

When there is ambiguity, the emphasis of annotation is on the author of the text's perspective. For example, in *"Samuel Pisar, an Auschwitz survivor said"*, the expression "Auschwitz survivor" only indicates who this person is, so we are not in the position of assigning a positive or negative connotation to *survivor* or *Auschwitz*. Instead, implied opinions should be annotated. For example, in *"Is China drinking our milkshake?"* there is a positive group "[our]$_M$ [milkshake]$_T$" and a negative one "[drinking]$_M$ [milkshake]$_T$", even if *drink* and *milkshake* do not express any sentiment by themselves.

5. RESULTS

So far, we have annotated 307 English sentences for a total of 6987 tokens with the annotation process still ongoing, and we plan to start with Italian and Russian soon. The original and annotated texts are available on-line[4]. In Table 1 more details are shown, including how many words are embedded in appraisal groups. This figure does not include duplicates, i.e. targets or modifiers repeated because used in more than one appraisal group, but it does include multi-word expressions. The number of modifiers and targets is not in proportion 2:1 to that of the appraisal groups because we mark them also in the case in which they can not be linked in a group.

We have found that 'appreciation' is the most common attitude across the domains, which is consistent with the fact that, in most cases, the target type to which the modifier attitude refers is 'thing'. This is followed by 'judgement' linked to type 'person'. Polarity has been 'marked' 42 times out of 901, thus indicating the times in which a negation has been encountered. Force, which is another important attribute of the modifier, has instead value 'reverse' 39 times and 'high' 62 times out of 901, compared to 'low' 7 times. Such data demonstrate how important is to mark reversals, as well as adverbs that increase or diminish the intensity of the modifiers. In Table 2 we show a detailed picture of the orientation types across domains.

We compared the contextual orientation manually annotated by us with the prior orientation included in the *NRC Word-Emotion Association Lexicon* [11], whose annotations were manually done through *Amazon's Mechanical Turk*, and the *Roget Thesaurus*[5] used as source for target terms. The lexicon has entries for about 24200 word–sense pairs, corresponding to 14200 word types. From the dictionary we considered only the values 'positive' and 'negative' for sentiment, although it also includes eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust).

All the words were first lemmatised and trimmed to avoid repetitions of singular and plural, as well as repetition of the same word with wrongly-annotated extra spaces. Our analysis was on 1872 words included in the appraisal groups. We calculated that words coming from the appraisal groups were present in the sentiment dictionary only 721 times, i.e. approximately 38.51% were covered.

We classified them in 4 major categories:

- Agreeing words: words whose dictionary orientation agrees with that of the appraisal group they are taken from.
- Disagreeing words: words whose dictionary orientation does not agree with that of the appraisal group they are taken from.
- Agreeing and disagreeing words: words whose dictionary orientation sometimes agrees and sometimes does not agree with that of the appraisal group they are taken from.
- Ambiguous words: words who already have both positive and negative values in the dictionary.

In Table 3, we show the number of times in which prior orientation and contextual orientation are agreeing, disagreeing and ambiguous for words taken from the appraisal groups and present in the sentiment dictionary.

Agreeing words cover the 70% of the total times words were found in the dictionary. This means that we can rely to a certain extent to the dictionary orientation, but not if we aim at more accuracy. The list includes reasonable out-of-context positive words (e.g., *almighty*, *friendly*, *respect*, *reward*, *grateful*), as well as out-of-context negative words (e.g., *abuse*, *adversary*, *crisis*, *destruction*, *failure*, *violence*).

Disagreeing words cover about the 28% of the total times words were found in the dictionary. This is important as it shows how crucial the context is. To further clarify the comparison, we also have both the prior and contextual orientation displayed, along with the appraisal groups in which those words appear. We found that words

[4] http://corpus.leeds.ac.uk/marilena/SentiML

[5] http://www.gutenberg.org/ebooks/10681

Domain	Appraisal groups	Targets	Modifiers	Sentences	Words	% annotated words
Political speeches	601	515	577	157	3783	27.54%
News	237	207	231	100	2281	19.33%
Talks	98	87	93	50	923	19.39%
TOTAL	936	809	901	307	6987	-

Table 1: Statistics on the annotated data

Orientation	Negative	Positive	Neutral	Ambiguous
Modifiers	170	236	149	346
Targets	173	147	328	164
Appraisal groups	363	556	2	15

Table 2: Orientation summary

Orientation	Frequency	Percentages
Agreeing words	509	70.60%
Disagreeing words	206	28.57%
Ambiguous words	6	0.83%

Table 3: Results of the comparison between prior and contextual orientation

with no sentiment values in the dictionary were included (e.g., *dark*, *rule*, *bear*, *prevent*, *change*), along with those with opposite sentiment. In this last category, we found that some had their orientation effectively influenced by the context (e.g., *real*/*crucial* issue, *hard* task) whereas some had a surprising orientation in the dictionary (e.g., *republic* and *government* were negative). We also found that many cases are of nouns linked by a preposition whose combination had reverse orientation (e.g., "infringement of liberty", "lack of freedom", "execution of citizen", "war/campaign against terrorism", "trade of sex", "prohibition of terrorism").

Agreeing and disagreeing words include the intersection of the two groups above, thus words such as *reversals* (e.g., *abolish*, *attack*, *oppose*, *question*), as well as words whose orientation is not difficult to imagine depending on what they refer to (e.g., *abandoned*, *absolute*, *afford*, *attention*, *choice*, *demand*, *depth*, *encourage*, *force*, *growth*, *important*, *interest*, *join*, *powerful*, *special*). Some words suggest an opposite out-of-context orientation (e.g., "innocent victims", "useless effort", "deserve to suffer"). In the case of some others that seem to be positive or negative *a priori* (e.g., *freedom*, *discrimination*, *liberty*, *peace*, *enemy*), we have found in fact that orientation depends on the context.

Ambiguous words were relatively rare, with only 4 (i.e. *influence*, *intervention*, *retirement*, *revolution*) being seen a total of 6 times.

Words not in dictionary represent 61,49% of the total times appraisal words appear. This is also an important figure to understand the limitations in the coverage of a sentiment dictionary, regardless on how good it is. While most of these are nouns, which reasonably might not be included in the sentiment dictionary, they are nevertheless still part of a group expressing an evaluation (e.g., *heritage*, *history*, *student*, *politics*). There are also cases, adjectives in particular, that should probably be included in a dictionary with prior orientation (e.g., *anti-terror*, *better*, *bitter*, *brave*, *wisely*, *weak*). For some of these, other forms are instead included in the dictionary (e.g., *bitterness*, *braveness*). The solution of checking words by their root to further increase the recall from the dictionary was excluded in order not to reduce precision, for example on phrasal verbs with different meaning (e.g., *get rid* vs *get*). Finally this category also includes multi-word expressions (e.g., "in dark", "in doubt", "in practice", "in search of", "in the face of", "in the light of day", "in the midst of"), as well as prepositions (e.g., *with*, *without*).

In terms of the speed of annotation, an expert annotator in our experiments was able to annotate 50 appraisal groups per hour on average, along with their targets and modifiers. On average each sentence (depending on the domain) contains 4 appraisal groups, which leads to performance of about 12 sentences per hour. Among the cases mentioned in Section 4, we have found that nouns linked by prepositions, followed by non-sentiment words and phrasal verbs are the most common.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have described an annotation scheme aimed at a comprehensive analysis of groups consisting of a modifier and target carrying sentiment. The manual annotation step is a time-consuming task, so we tried to keep it as simple as possible without losing important linguistic features, such as both the prior and contextual orientation of words. We have also demonstrated that prior orientation given in the dictionary is different from the correct one given by the context in 28% of the cases, and that, in general, the dictionary has a relatively low coverage.

Future work will include more investigations of the inter-annotator agreement which, on the basis of previous studies [18], should not be difficult to obtain on the difference in orientation (positive/negative), but it is likely to be more challenging on such linguistic features as appraisal types and force values.

In addition, we are in the process of developing a pipeline that is able to automatically extract appraisal groups by using dependency parsing. The work will be conducted in English, Italian and Russian.

7. ACKNOWLEDGMENTS

We take the chance to thank the reviewers for their precious feedback. The first author would also like to thank Michele Filannino for his useful insights and support.

8. REFERENCES

[1] S. Argamon, K. Bloom, A. Esuli, and F. Sebastiani. Automatically Determining Attitude Type and Force for Sentiment Analysis. In *Proceedings of the 3rd Language and Technology Conference (LTC'07)*, pages 369–373, Poznan, Poland, 2007.

[2] K. Bloom, N. Garg, and S. Argamon. Extracting appraisal expressions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of*

the Main Conference, pages 308–315, Rochester, New York, Apr. 2007.

[3] M. Cettolo, C. Girardi, and M. Federico. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May 2012.

[4] P. Dodds and C. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, August 2010.

[5] S. B. Dyson. Text annotation and the cognitive architecture of political leaders: British prime ministers from 1945-2008. *Journal of Information Technology & Politics*, 5(1):7–18, 2008.

[6] M. A. Halliday. *An Introduction to Systemic Functional Linguistics*. London:Arnold, 2 edition, 1994.

[7] N. Ide and L. Romary. International standard for a linguistic annotation framework. *Nat. Lang. Eng.*, 10(3-4):211–225, Sept. 2004.

[8] R. Johansson and A. Moschitti. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3), 2013.

[9] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[10] J. R. Martin and P. R. White. *The language of evaluation*. Palgrave Macmillan, Basingstoke and New York, 2005.

[11] S. Mohammad. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, June 2011.

[12] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 806–814, Stroudsburg, PA, USA, 2010.

[13] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[14] J. Read, D. Hope, and J. Carroll. Annotating expressions of appraisal in English. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 93–100, Stroudsburg, PA, USA, 2007.

[15] A. Stubbs. Mae and mai: Lightweight annotation and adjudication tools. In *Linguistic Annotation Workshop*, pages 129–133, 2011.

[16] M. Taboada and J. Grieve. Analyzing appraisal automatically. In *In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161, 2004.

[17] C. Whitelaw, N. Garg, and S. Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 625–631, New York, NY, USA, 2005.

[18] T. A. Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh, 2008.

[19] Y. Zhao, B. Qin, W. Che, and T. Liu. Appraisal expression recognition with syntactic path for sentence sentiment classification. *Int. J. Comput. Proc. Oriental Lang.*, 23(1):21–37, 2011.