Multiple views as aid to linguistic annotation error analysis

Marilena Di Bari
University of Leeds
mlmdb@leeds.ac.uk

Serge Sharoff
University of Leeds
s.sharoff@leeds.ac.uk

Martin Thomas
University of Leeds
m.thomas@leeds.ac.uk

Abstract

This paper describes a methodology for supporting the task of annotating sentiment in natural language by detecting borderline cases and inconsistencies. Inspired by the co-training strategy, a number of machine learning models are trained on different *views* of the same data. The predictions obtained by these models are then automatically compared in order to bring to light highly uncertain annotations and systematic mistakes. We tested the methodology against an English corpus annotated according to a fine-grained sentiment analysis annotation schema (SentiML). We detected that 153 instances (35%) classified differently from the gold standard were acceptable and further 69 instances (16%) suggested that the gold standard should have been improved.

1 Introduction

This work pertains to the phase of testing the reliability of human annotation. The strength of our approach relies on the fact that we use multiple supervised machine learning classifiers and analyse their predictions in parallel to automatically identify disagreements. Those, in fact, ultimately lead to the discovery of borderline cases in the annotation, an expensive task in terms of time when carried out manually.

Predictions with a number of different labels are manually analysed, since they may indicate inconsistencies in the annotation and cases difficult to annotate. Conversely, cases with high agreement suggest that the annotation schema is reliable. On the one hand, the analysis of those disagreements, in conjunction with the gold annotations, provides fresh insights about the efficacy of the features provided to the classifiers for the learning phase. On the other hand, when all the classifiers agree on a wrong annotation, it is a strong signal of ambiguity in the annotation schema and/or guidelines.

In Section 2 we briefly introduce the data to which we apply the methodology described in Section 3. In Section 4 we report results. In Section 5 we mention studies related to ours and in Section 6 we draw conclusions and identify steps for future work.

2 Data

We tested our methodology on the *SentiML* corpus (Di Bari et al., 2013) for which the annotation guidelines, as well as the original and annotated texts, are publicly available [1]. The corpus consists of 307 English sentences (6987 tokens), taken from political speeches, *TED* talks (Cettolo et al., 2012), and news items from the *MPQA opinion corpus* (Wilson, 2008).

The aim of its annotation is to encapsulate opinions in pairs, by marking the role that each word takes (modifier or target). For example, in

> "More of you have lost your homes and even more are watching your home values plummet"

there would be two pairs: *modifier* "lost" and *target* "homes", and *modifier* "values" and *target* "plummet". Such two pairs are called *appraisal groups*.

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

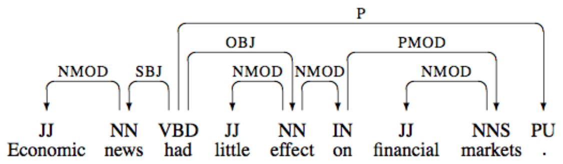

Figure 1: Example of dependency tree. Dependency trees provide features for the machine learning step.

For each of these elements several features are annotated that are believed to improve the task of sentiment analysis. The study presented here relates to the automatic identification of modifiers and targets.

3 Methodology

To test our methodology we selected a corpus for which various types of linguistic information related to appraisal groups were annotated. We started with the identification of modifiers and targets, since this represents the base of all the other levels of annotation.

To test the reliability of annotation we set 10% of our annotated corpus aside, and performed the machine learning part of the study on the remaining 90% of our corpus.

The first step consists of preparing the features for the machine learning phase. The optimal set to model the annotation task varies from problem to problem. We used the following:

- Word features, representing the ordinal identifier, word form, lemma and POS tag of each word.
- Contextual features, representing the lemma and POS tags of the preceding and succeeding words.
- Dependency-based features, representing the reference to the word on which the current token depends in the dependency tree (*head*) along with its lemma, POS tag and relation type (see Figure 1) (Nivre, 2005).
- Number of linked modifiers, representing the number of adjectives and adverbs linked to the current word in the dependency tree.
- Role, representing the predicted role (modifier or target) of the current token in conveying sentiment. The predictions are computed using fixed syntactic rules.
- Gazetteer-based sentiment. We used the *NRC Word-Emotion Association Lexicon* (Mohammad, 2011) to represent the *a-priori* sentiment of each word, i.e. regardless of its context.

Once the features are ready, two or more feature partitions (called *views* in the co-training strategy) have to be defined in order to be as orthogonal as possible (Abney, 2007). We opted for a linguistically-grounded dichotomy: lexical features (word features, role and gazetteer-based sentiment) versus syntactic features (contextual and dependency-based features, number of linked modifiers). The training and test sets are split accordingly.

At this point, machine learning classifiers are chosen. These need to be confidence-rated, i.e. able to provide a confidence rate for each prediction. In our experiments we selected Naïve Bayes, Radial Basis Function Network and Logistic Regression[2]. These models rely on very different strategies, which makes the analysis more reliable. We discarded Support Vector Machines since in our preliminary experiments they achieved high precision (a range between 0.60 and 0.77 across modifiers and targets), but very low recall (a range between 0.05 and 0.06 across modifiers and targets), which resulted in a very low F-measure (a range between 0.09 and 0.11 across modifiers and targets).

A model for each combination of view and classifier is then produced and tested on the test set. We performed a 10-fold cross-validation. In the test phase, we opted for a numerical threshold of 0.67 to consider the predictions reliable. A prediction with a confidence lower than the threshold is considered uncertain.

For each instance we obtained six predictions, which potentially differ from one another. The agreement score is calculated for each class in order to identify the most frequent prediction.

[2] In each case we used the implementation provided by WEKA (`http://www.cs.waikato.ac.nz/~ml/weka/`).

Feature set	**Classifier**	**Modifier**			**Target**		
		Precision	**Recall**	$\mathbf{F}_{\beta=1}$	**Precision**	**Recall**	$\mathbf{F}_{\beta=1}$
Lexical	Naïve Bayes	**0.71**	0.10	0.48	**0.82**	0.12	0.43
	RBF Network	0.52	**0.56**	**0.54**	0.51	**0.59**	**0.55**
	Logistic regression	0.59	0.42	0.49	0.61	0.48	0.54
Syntactic	Naïve Bayes	0.46	0.48	0.47	**0.82**	0.12	0.43
	RBF Network	0.49	0.35	0.40	0.55	0.50	0.53
	Logistic regression	0.58	0.22	0.32	0.60	0.41	0.49

Table 1: Performance of the classifiers trained on two views, lexical and syntactic. Experiments have been performed using 10-fold cross-validation.

At this point, only the predictions different from the gold annotations are considered: the higher the agreement score, the more the instance is interesting in the context of our analysis.

The final step consists of manually investigating such cases to shed light on the errors. In this experiment we opted for the use of a simple protocol based on the following classification schema:

- W (wrong), where the classifiers disagree with the gold annotation, which we judge to be correct.
- A (ambiguous), where the classifiers disagree with the gold annotation and we judge both to be valid. In such cases, the guidelines need to be clearer or the annotation method could have been simpler.
- M (to modify), where we judge that the gold annotation is incorrect.

This approach has the advantage of yielding a much reduced subset of instances to be examined manually, with respect to the full set.

4 Results

Table 1 shows the performances of the six models obtained from the training of each combination of view and classifier, mentioned in Section 3. F-measures for modifiers range between 0.32 and 0.54 for modifiers, and 0.43 and 0.55 for targets. Overall, the RBF Network trained on the lexical view performs best. However, there is no huge difference in general in performances between the lexical and the syntactic feature sets, which is good in the light of data sparseness.

Performance on the the empty class (no category assigned) was exceptionally good, as 76% was predicted out of the gold 77%, whereas the performance on the modifiers was 4% out of the gold 12% and the performance on the targets was 5% out of the gold 11%. Although the annotation allows each token to be simultaneously annotated as modifier and target, we have not reported the performances for the MT class as the cases were not significant. Finally, there was a 15% of cases in which the classifiers were not confident.

In relation to the manual classification of errors (see final paragraph of Section 3) we found that, out of the total test instances (2066), in 436 cases the most predicted class differed from the gold standard: the label *W* was assigned 214 times (49%), the label *A* was assigned 153 times (35%), the label *M* was assigned 69 times (16%). *W* was mostly assigned when the modifier or the target was correctly identified, but not its counterpart in the pair (e.g., "way forward", "blame society", "wrong side"). It was also assigned when a word was correctly identified as evoking sentiment (e.g., "destroy", "flourish"), but only the first of two or more targets was identified (e.g., "women and children", "the city and the country").

A was assigned when an adverb was annotated as modifier (e.g., "through corruption", "seize gladly", "tragically reminded"): these are cases in which human annotators decide to include the adverb if it is regarded as important for the sentiment. Other cases in which the label has been used is with compound modifiers (e.g., "face to face", "in the face of"), phrasal verbs (e.g., "turn back", "carried forth", "came forth") and difficult couples to link (e.g., "instruments with which we meet them" [challenges]). Finally, this label was also used in cases in which the prediction was sensible, but considered less accurate than the gold one (e.g., in "enjoy relative plenty", the gold standard was "enjoy plenty" and the classifiers predicted “relative plenty”).

M was assigned when another modifier had been wrongly annotated by the annotator, instead of modifying the value of the force of the current one (e.g., in “much more”, only “more”' should have been annotated with *high* force), in the case of couples with no sentiment (e.g., “future generations”, “different form”), of couples not previously identified (e.g., “stairway filled with smoke”, “icy river”) or couples that could have been annotated in an easier way (e.g., “provoke us to step up and do something”, “image resonates with us”).

5 Related work

Evaluating the reliability of human annotation is a challenging and widely studied task (Pustejovsky and Stubbs, 2012). The standard solution is the measurement of an inter-annotator agreement (IAA) coefficient according to a variety of formulae that depend on the characteristics of the annotation setting (Artstein and Poesio, 2008).

For example, in the case of Wilson (2008) and Read and Carroll (2007), it was useful to understand inconsistencies in the selection of the span for attitudes and targets. Since this represents only one of the commonly recognized challenges, some studies have focused on practically testing a methodological framework for schema development for fine-grained and quality semantic annotations. (Bayerl et al., 2003).

Our approach varies from the standard procedure in ways similar to that of Snow et al. (2008). For each expert annotator (six in total) they trained a system using only the judgements provided by these annotators, and then created a test set using the average of the responses of the remaining five labellers on that set. This resulted in six independent expert-trained systems. The difference with our methodology is that we trained six independent classifiers, but based on judgements of only one human annotator, and compared the average of the responses of six classifiers with the gold standard.

Jin et al. (2009) also used the strategy of selecting the labelled sentences agreed upon by their classifiers and achieved good performances in the task of identifying opinion sentences.

Finally, our methodology is also similar to one of those mentioned by Yu (2014). The author used the traditional co-training strategy, i.e. providing a small pool of unlabelled data to two classifiers with confidence rates, in order to obtain automatically labelled examples that would be added to an initial set of labelled ones. Subsequently, this final large set is used to train the the two classifiers and a combination of them (constructed by multiplying their predictions) is eventually the one used to label new documents. Five strategies were applied to obtain the views: (a) using unigrams and bigrams as features, (b) randomly splitting the feature set in two, (c) using two different supervised learning algorithms because they would provide useful examples to each other since based on different learning assumptions; (d) randomly splitting the training set, and (e) applying a character-based language model (CLM) and a bag-of-words model (BOW). We extended the third strategy by using three classifiers and two different views for each of them, and by applying this to the task of annotation validation rather than semi-supervised learning.

6 Conclusions

In this paper we have presented a methodology that makes use of multiple classifiers (based on different views) in order to detect inconsistent annotations and borderline cases. In our test set, we found that in 35% of the wrongly classified cases the predictions were different but acceptable, and in the 16% of them the predictions suggested that the gold standard was wrong. On the other hand, the data resulting from such procedure related to non-disagreeing predictions can be regarded as expression of either the efficacy of the annotation schema and guidelines or the features used for the machine learning step.

Our next goal is to improve the performances of the classifiers over the instances that were incorrectly handled, currently accounting for the 26% in our test set. We will also test the same methodology over the extraction of the link between targets and modifiers (appraisal groups). The machine learning models, the datasets and the error analysis are publicly available in order to ensure reproducibility [3].

[3] http://corpus.leeds.ac.uk/marilena/SentiML/LAW2014_error_analysis.zip

References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC, 1st edition.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Petra S. Bayerl, Harald Lüngen, Ulrike Gut, and Karsten I. Paul. 2003. Methodology for reliable schema development and evaluation of manual annotations. In *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003*, pages 17–23.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. 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.

Marilena Di Bari, Serge Sharoff, and Martin Thomas. 2013. SentiML: Functional annotation for multilingual sentiment analysis. In *DH-case 2013: Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities*, ACM International Conference Proceedings.

Wei Jin, Hung H. Ho, and Rohini K. Srihari. 2009. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1195–1204, New York, NY, USA. ACM.

Saif Mohammad. 2011. 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.

Joakim Nivre. 2005. Dependency grammar and dependency parsing. Technical report, Växjö University.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. Oreilly and Associate Series. O'Reilly Media, Incorporated.

Jonathon Read, David Hope, and John Carroll. 2007. Annotating expressions of appraisal in English. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 93–100, Stroudsburg, PA, USA.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh.

Ning Yu. 2014. Exploring co-training strategies for opinion detection. *Journal of the Asssociation for Information Science and Technology*.