Entities' Sentiment Relevance

Zvi Ben-Ami
The Hebrew University
Jerusalem, ISRAEL
zvi.benami@mail.huji.ac.il

Ronen Feldman
The Hebrew University
Jerusalem, ISRAEL
ronen.feldman@huji.ac.il

Binyamin Rosenfeld
Digital Trowel
New York, USA
grurgrur@gmail.com

Abstract

Sentiment relevance detection problems occur when there is a sentiment expression in a text, and there is the question of whether or not the expression is related to a given entity or, more generally, to a given situation. The paper discusses variants of the problem, and shows that it is distinct from other somewhat similar problems occurring in the field of sentiment analysis and opinion mining. We experimentally demonstrate that using the information about relevancy significantly affects the final sentiment evaluation of the entities. We then compare a set of different algorithms for solving the relevance detection problem. The most accurate results are achieved by algorithms that use certain document-level information about the target entities. We show that this information can be accurately extracted using supervised classification methods.

1 Introduction

Sentiment extraction by modern sentiment analysis (SA) systems is usually based on searching the input text for sentiment-bearing words and expressions, either general (language-wide) or domain-specific. In most common SA approaches, each such expression carries a polarity value ("positive" or "negative") which is possibly weighted. The sum of all polarity values from all expressions found in a text becomes the sentiment score for the whole text.

People are, however, usually interested in sentiments regarding some entity or situation, and not in sentiments of a particular document. A natural way to make the SA more focused is to explicitly bind each sentiment expression to a specific entity, or to a small set of entities from among all entities mentioned in the document. The choice of which entity to bind a sentiment expression to, can be made according to the proximity (physical, syntactical, and/or semantic) and/or salience of the entities.

In this paper, we argue that all of these methods can be useful in different contexts, and so the best single algorithm should use all available proximity information, of all kinds, together with additional context information –position in the document, section, or paragraph; proximity of other entities; lexical contents; etc. One of the most important context information is the type of relation between the target entity and the document – whether the entity is the main topic of the document, or one of several main topics, or mentioned in passing, etc.

Another layer that we'd like to add concerns the interaction of different entity types during SA. In a typical situation, there is only one entity type which is the target for SA. In such cases, clearly distinguishing between the relevancy of target and non-target entities types is not essential. For example, when the general topic is a COMPANY, and there is a sentiment expression referring to a PERSON or a PRODUCT, this sentiment expression is still relevant to the company and can be regarded as such. In other situations, SA users may be specifically interested in an interaction between entities of different types. For example, in a medical forum setting, it may be interesting to know the users' sentiments regarding a given DRUG in the context of a given DISEASE. We will show that such situations are modeled well enough using intersections of regions of relevance of the participating entity types, with the relevance region for each type calculated separately.

We purposefully exclude possible interactions between entities of the same type, because they behave in a different way. The precise analysis of such interactions is a different topic from rele-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 87–92,
Baltimore, Maryland, USA, June 23-25 2014.

vance detection, and so it is mostly ignored in this paper.

2 Related Work

The task of SA has drawn the attention of many researchers worldwide (Connor et al., 2010; Liu, 2012; Loughran and Mcdonald, 2010; Pang and Lee, 2004; Turney, 2002). While most SA research is focused on discovering and classifying the expressions, some are also concerned with the targets of the expressions and explicitly identify the syntactic targets of sentiment expressions (Pang and Lee, 2004).

Other related works belong to the Passage Retrieval field, since the relevance detection problem can be construed as a specific form of passage retrieval problem (Liu and Croft, 2002; Tiedemann and Mur, 2008). Different approaches were suggested for passage retrieval (Buscaldi et al., 2010; Comas et al., 2012; Hearst, 1997; Lafferty et al., 2001; Lin et al., 2012; Liu and Croft, 2002; Lloret et al., 2012; O'Connor et al., 2013; Otterbacher et al., 2009; Salton et al., 1993; Wachsmuth, 2013), some are more sophisticated than others.

The closest approach to ours is the one of Scheible and Schütze (2013), but in contrast to them, we strive to discover sentiments' relevance for all entities (of a given type) mentioned in the document, not necessarily topical.

3 Entity Relevance

An instance of the sentiment relevance detection problem for a single entity consists of a text document, a sentiment expression within the document, and a target entity. The task is a binary decision: 'relevant' vs. 'irrelevant'. To solve this task, we can use any information that can be found by analyzing the document. Thus, we can assume that we know the parse trees of all sentences and the locations of all references of all entities in the document, including co-references.

In addition, we make use of an extra piece of information for each target entity – its "status within the document", or "document type with respect to the entity". We distinguish between several types which are intuitively clearly different:

- **'Target'** – the entity is the main topic of the document;
- **'Accidental'** – the entity is not the main topic of the document, and is mentioned in passing;
- **'RelationTarget'** – the main topic of the document is a relation between the entity and some other entities of the same type;
- **'ListTarget'** – the entity is one of a few equally important topics, dealt with sequentially.

In the datasets we use for experiments, each entity is manually annotated with its status within the document, which allows us to directly observe the influence of this data on the accuracy of relevance discernment. We also show that this data can be automatically extracted using supervised classification.

Since this paper is primarily a study of sentiment relevance, the actual sentiment expressions are not always labeled in our datasets. Instead, relevance ranges are annotated for each entity, in the style of passage retrieval problems, with the expectation that sentiment expressions relevant to an entity only appear in the parts of the document that are labeled as "relevant", and conversely, that all expressions appearing in parts labeled "irrelevant" are irrelevant. This way of annotating allows the comparing of different relevance detection strategies independently of the main sentiment extraction tool.

All of the algorithms discussed in this paper use the same document processing methods, thus allowing us to compare the algorithms themselves independent of the quality and specifics of the underlying NLP.

The multiple-entity relevance problem is distinguished from the single-entity relevance problem by the requirement for the sentiment expression to be relevant to several entities of different types. The problem is close to Relation Extraction in this sense. The examples we are interested in are in the medical domain and deal with three main entity types: PERSON, DRUG, and DISEASE, where PERSON is restricted to known physicians. While each of the entity types can be the target of a sentiment expression, the more interesting questions in this domain involve multiple entities, specifically, DRUG + DISEASE ("how effective is this drug for this disease?"), and PERSON + DRUG + DISEASE ("what does this physician say about using this drug to cure this disease?").

We solve the multiple-entity relevance problem by intersecting the relevance ranges of different-type entities, thus reducing the problem to the single-entity relevance detection. As such, the experiments regarding the multiple-entity relevance need only check the accuracy of this reduction. In the medical domain, at least, this accuracy appears to be adequate.

4 Relevance Algorithms

Each algorithm receives, as input, the text of the document, with labeled reference of the target entity and other entities of the same type. The labeled references also include all coreferential references, extracted automatically by an NLP system. The input text also includes labeled candidate sentiment expressions, either manually labeled or automatically extracted by a relevance-ignoring SA system[1]. The task of the algorithms is to label each candidate expression as relevant or irrelevant to the target entity. The algorithms are evaluated according to the accuracy (recall, precision, and F1) of this labeling of individual sentiment expressions.

This method produces a reasonably well-understandable quality measure (the percentage of expressions that the algorithms get right or wrong), and also allows us to compare algorithms focused on individual expressions and algorithms working on text ranges. The algorithms we evaluate are as follows:

- **Baseline** - Every expression is declared relevant. This is the standard mode of operation of document-level SA tools, although it is usually only applied to the 'Target' entities – the main topic(s) of the document.
- **Physical-proximity-based** - A text-range focused algorithm, which labels pieces of text as relevant or irrelevant according to their placement relative to the references of the target entity and other entities of the same type, as well as some other contextual clues, such as paragraph boundaries. Generally, the mentioning of an entity starts its relevance range (and stops the relevance range of the previously mentioned entity). For the first entity reference in a paragraph, the range also extends backward to the beginning of the sentence. There are three flavors of the algorithm, specifically adapted for different document-types-with-respect-to-the-target-entity:
  - '**Proximity-Accidental**' - stops relevance ranges at paragraph boundaries,
  - '**Proximity-Targeted**' - restarts relevance ranges at paragraph boundaries (every paragraph is assumed relevant at the start, unless another entity is mentioned).
  - '**Proximity-List**' - interpolates relevance ranges over intermission paragraphs, unless they are explicitly irrelevant (e.g., containing references of other entities of the same type).
- **Syntactic-proximity-based** - An expression-focused algorithm, which labels expressions as relevant or irrelevant according to their distance to various entity references in the dependency parse graph. There are two flavors of the algorithm: direct and reverse. The former considers an expression relevant only if it is closest to the target entity from among all entities of the same type, and the distance is sufficiently close. The latter considers an expression irrelevant only if it has the above-described relation to some non-target entity of the same type. The rationale for the two flavors is the distinction between 'Targeted' and 'Accidental' document types regarding the target entity. For the 'Accidental' entities, a sentiment expression is assumed to be relevant only if it is explicitly connected to the entity. For 'Targeted' entities, an expression is irrelevant only if it is explicitly connected to some other entity of the same type.
- **Classification-based** - This algorithm considers each candidate sentiment expression as an instance of a binary classification problem, to be solved using supervised classification. For evaluating this algorithm, some part of the test corpus is used for training, and the other for testing, with N-fold cross-validation. The features for classification may use any information present in the input.

  In the current experiments, we use references of target and non-target entities, appearances of paragraph and document boundaries, length of syntactic connections to target and non-target entities, when available, and explicit entity status within documents, when available. The (binary) classification features are built from sequences of up to 5 occurrences of the above-described pieces, with the pieces appearing before and after the sentiment expression tracked separately. For classification, we use a linear classifier with Large Margin training (regularized perceptron, as discussed in Scheible and Schütze, (2013)).
- **Sequence-classification-based** - The algorithm uses exactly the same features as the direct classification-based above, but instead of considering each expression separately, it con-

[1] In our experiments, we also use a standalone automatic Financial SA system from Feldman et al. (2010), working in the 'ignore relevance' mode, which (1) finds and labels all entities of the target type(s); (2) resolves all coreferences for the target entity type(s); (3) finds and labels all sentiment expressions, regardless of their relevance; and (4) provides dependency parses for all sentences in the corpus.

siders them as a sequence, one per document. So, instead of a Large Margin binary classifier, a probabilistic sequence classifier is used (CRF, as discussed in Lafferty et al. (2001)).

5 Experiments

For the experiments, we use two manually-annotated corpora[2], a financial corpus[3] and a medical[4] corpus. In the Financial corpus, COMPANIEs are used as target entities and in the medical corpus, DISEASEs, DRUGs and PERSONs are the entity types that are used as target entities. For the purpose of the experiments, we are interested only in single-entity sentiments about DRUGs, and multiple-entity sentiments about DRUGs + DISEASEs, and DRUGs + DISEASEs + PERSONs.

The evaluation metrics in all of the experiments are precision, recall, and F1. For the classification-based algorithms, unless stated otherwise, we use 10-fold cross-validation.

5.1 Experiment: Importance of relevance

In the first experiment, we demonstrate the importance of using relevance when calculating the consolidated sentiment score of an entity within a set of documents. For each entity, we set the 'correct' consolidated sentiment score to the average of polarities of all sentiments in a corpus which are labeled as relevant to the entity. Then, we compare the correct value to the two scores calculated without considering relevance:

- **'Baseline'** - the average of polarities of all sentiments in all documents where the entity is mentioned, and
- **'TargetedOnly'** - the average of polarities of all sentiments in the documents where the entity is labeled as target (main topic of the document). This case models the typical state of a relevance-agnostic SA system.

For this evaluation, we only compare the sign of the final sentiment scores, without considering their magnitudes (unless it is close to zero, in which it is considered 'neutral'). The errors at this level indicate definite SA errors – miscalculating entity's sentiment into its opposite.

The results of the evaluation are as follows: The 'Baseline' scores show a large difference from the correct scores, with 33% and 38% of entities having wrong final polarity in the financial (COMPANY) and medical (DRUG) domains, respectively. The 'TargetedOnly' scores are somewhat closer to correct, with 12% and 28% of entities with incorrect final polarities. However, the 'TargetedOnly' method naturally suffers from a very low recall, with only 19% and 38% of entities covered in the financial and medical domains, respectively.

5.2 Experiment: Influence of entity status

In this experiment, we compare the performance of various algorithms while either providing or withholding the information about the document-type-with-respect-to-the-target-entity.

The performance of the physical proximity algorithms on the financial corpus is shown at the top left hand side of Table 1. The set of all instances of relevance detection problems in the corpus (an instance consists of a sentiment expression within a text, together with a target entity) is divided into three subsets, according to the status of the target entity within the document. As expected, the three flavors of the physical proximity algorithm perform much better on the corpus subsets they are adapted to. At the bottom left hand side of Table 1, we similarly show the performance of the two flavors of the syntax-proximity-based algorithm on the medical domain (DRUG entities). Same as above, there is a large difference in the performance of the two flavors of the algorithm on different subsets of the problem set. Finally, at the top of Table 2, we compare the performance of the two classification-based algorithms on the two (whole) problem sets, while either keeping or withholding the entity status information from the classifier. The difference in results is less pronounced here, but is still noticeable. The reason for the smaller difference, we hypothesize, is the ability of the classifiers to partially infer the entity status from the various context clues that are used as classification features (see the experiment 5.3).

5.3 Experiment: Automatic identification of entity status using classification.

In this experiment, we confirm that it is possible to identify the entity status within documents using supervised classification.

[2] Fully annotating texts for semantic relevance is an arduous task, thus the used annotated corpora are relatively small. Sample can be found at http://goo.gl/6HONHP.

[3] A corpus of 160 financial news documents on at least one entity of interest, of average size ~5Kb, downloaded from various financial news websites. The dataset mentions 424 different companies.

[4] A corpus of 160 documents, of average size ~7Kb, downloaded following Google queries on a set of a few common drugs and diseases. The dataset mentions 722 different people, 46 diseases, and 175 drugs.

	Experiment 5.2 (Precision/Recall/F1)				Experiment 5.3 (F1, (diff. in F1 from exp. 5.2)			
	Accidental	Targeted	List	Whole	Accidental	Targeted	List	Whole
Proximity-Accidental	84/43/**57**	93/76/84	92/74/82	92/72/81	60 (+2.6)	79 (-5.5)	83 (+1.1)	
Proximity-Targeted	31/50/38	90/**84**/87	55/89/68	63/83/72	38 (-0.4)	82 (-5.2)	73 (+4.3)	
Proximity-List	58/44/50	90/83/87	88/83/**86**	85/80/82	52 (+2.1)	81 (-5.9)	87 (+1.6)	
Proximity-Combined				89/80/**84**				83 (-1.2)
Syntactic-Prox.-Direct	93/48/**64**	99/42/60			65 (+0.8)	59 (-0.2)		
Syntactic-Prox.-Inverse	04/72/08	70/66/**68**			8 (-0.2)	76 (+6.4)		

Table 1. Performance of different algorithms on three subsets of the corpus with a different status of the target entity within the document.

Experiment	Algorithm	Financial	Medical
Experiment 5.2 (Prec./ Rec,/F1).	Classification (with entity status info)	90/86/**88**	84/88/**86**
	Classification (without entity status info)	89/85/87	87/81/84
	Sequence Classification (with entity status info)	96/84/**90**	99/84/**91**
	Sequence Classification (without entity status info)	96/83/89	95/85/90
Experiment 5.3 (F1, (diff. in F1 from exp. 5.2))	Classification	**86.7** (-0.9)	83.9 (-2.0)
	Sequence Classification	**89.7** (+0.1)	90.9 (-0.3)
Experiment 5.5 (F1)	Baseline	37.2	28.6
	Physical Proximity	84.1	79.5
	Syntactic-Proximity	43.8	54.6
	Classification	87.6	85.9
	Sequence-Classification	**91.2**	**89.6**

Table 2. Performance of different algorithms on the different domains.

The results of direct evaluation show that the accuracies of the Medical and Financial corpora (using 10-fold X-validation) are 87.8% and 82.2% respectively, and the accuracy when using the Medical corpus for training the Financial corpus for testing and vice versa, are 78.2% and 86.1% , respectively.

The results of relevance detection using the automatically extracted entity status values are shown at the right hand side of Table 1 and in the middle of Table 2, which utilize the same datasets and algorithms as at the left hand side of Table 1 and at the top of Table 2. As can be seen from the tables, the drop in performance is small, demonstrating the success of classification-based extraction of entity status information.

5.4 Experiment: Cross-domain applicability

In this experiment, we test how well the classifiers trained on data from one domain work on input from a different domain.

The classification results using different types of training data are shown in Table 3.

	Classification	Sequence classification
Medical 2-fold/10-fold	84.6/85.9	85.7/89.6
Train on Fin, test on Med	83.5	86.8
Financial 2-fold/10-fold	86.1/87.6	90.3/91.2
Train on Med, test on Fin	85.4	91.0

Table 3. Performance of classification-based algorithms using different training data (F_1).

The table confirms general independence of the classification performance on the domain. Comparing the 2-fold and 10-fold cross-validation results (the difference is equivalent to doubling the amount of training data), shows that the amount of training data is sufficient.

5.5 Experiment: Overall performance of algorithms

In this experiment, we simply compare the overall accuracy of various algorithms for relevance discernment, operating at their best parameters. The results are shown at the bottom of Table 2. Overall, classification-based algorithms perform better than the deterministic ones, with sequence-classification performing significantly better than direct classification. Syntactic proximity-based is precise, but has relatively low recall, reducing its overall performance. Physical proximity-based is simplest, and produce reasonably high overall results, although worse than the best-performing classification-based methods.

6 Conclusion

The results are mostly intuitively understood and confirm the expectations. We confirmed that relevance detection is essential for producing correct consolidated SA results. We found that the entity status within the document is one of the important clues for solving the relevance detection problem, and showed that this information can be effectively automatically extracted using supervised classification. We also compared several algorithms for relevance detection, with the results that classification-based algorithms generally outperform simpler ones based on the same clues, although a very simple proximity-based algorithm performs reasonably well if allowed to use the entity status information.

Acknowledgments

This work is supported by the Israel Ministry of Science and Technology Center of Knowledge in Machine Learning and Artificial Intelligence and the Israel Ministry of Defense.

References

Buscaldi, D., Rosso, P., Gómez-Soriano, J., Sanchis, E., 2010. Answering questions with an n-gram based passage retrieval engine. J. Intell. Inf. Syst. 34, 113–134. doi:10.1007/s10844-009-0082-y

Comas, P.R., Turmo, J., Màrquez, L., 2012. Sibyl, a factoid question-answering system for spoken documents. ACM Trans. Inf. Syst. 30, 19:1–19:40. doi:10.1145/2328967.2328972

Connor, B.O., Balasubramanyan, R., Routledge, B.R., Smith, N.A., 2010. From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series, in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 122–129.

Feldman, R., Rosenfeld, B., Bar-haim, R., Fresko, M., 2010. The Stock Sonar — Sentiment Analysis of Stocks Based on a Hybrid Approach, in: Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference. pp. 1642–1647.

Hearst, M.A., 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist. 23, 33–64.

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data., in: Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001).

Lin, H.-T., Chi, N.-W., Hsieh, S.-H., 2012. A concept-based information retrieval approach for engineering domain-specific technical documents. Adv. Eng. Informatics 26, 349–360. doi:http://dx.doi.org/10.1016/j.aei.2011.12.003

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

Liu, X., Croft, W.B., 2002. Passage retrieval based on language models, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM '02. ACM, New York, NY, USA, pp. 375–382. doi:10.1145/584792.584854

Lloret, E., Balahur, A., Gómez, J., Montoyo, A., Palomar, M., 2012. Towards a unified framework for opinion retrieval, mining and summarization. J. Intell. Inf. Syst. 39, 711–747. doi:10.1007/s10844-012-0209-4

Loughran, T.I.M., Mcdonald, B., 2010. When is a Liability not a Liability ? Textual Analysis , Dictionaries , and 10-Ks Journal of Finance , forthcoming. J. Finance 66, 35–65.

O'Connor, B., Stewart, B.M., Smith, N.A., 2013. Learning to Extract International Relations from Political Context, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Sofia, Bulgaria, pp. 1094–1104.

Otterbacher, J., Erkan, G., Radev, D.R., 2009. Biased LexRank: Passage retrieval using random walks with question-based priors. Inf. Process. Manag. 45, 42–54. doi:http://dx.doi.org/10.1016/j.ipm.2008.06.004

Pang, B., Lee, L., 2004. A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.

Salton, G., Allan, J., Buckley, C., 1993. Approaches to passage retrieval in full text information systems, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93. ACM, New York, NY, USA, pp. 49–58. doi:10.1145/160688.160693

Scheible, C., Schütze, H., 2013. Sentiment Relevance, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Sofia, Bulgaria, pp. 954–963.

Tiedemann, J., Mur, J., 2008. Simple is best: experiments with different document segmentation strategies for passage retrieval, in: Coling 2008: Proceedings of the 2nd Workshop on Information Retrieval for Question Answering, IRQA '08. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 17–25.

Turney, P., 2002. Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews, in: Proceedings of the Association for Computational Linguistics (ACL). pp. 417–424.

Wachsmuth, H., 2013. Information Extraction as a Filtering Task Categories and Subject Descriptors, in: To Appear in Proc. of the 22th ACM CIKM.