Sentiment-Stance-Specificity (SSS) Dataset: Identifying Support-based Entailment among Opinions.

Pavithra Rajendran, Danushka Bollegala, Simon Parsons
University of Liverpool, University of Liverpool, Kings College London
Pavithra.Rajendran@liverpool.ac.uk, danushka.bollegala@liverpool.ac.uk, simon.parsons@kcl.ac.uk

Abstract

Computational argumentation aims to model arguments as a set of premises that either support each other or collectively support a conclusion. We prepare three datasets of text-hypothesis pairs with support-based entailment based on opinions present in hotel reviews using a distant supervision approach. Support-based entailment is defined as a specific opinion (premise) that supports as well as entails a more general opinion and together support a generalized conclusion. A set of rules is proposed based on three different components — *sentiment*, *stance* and *specificity* to automatically predict the support-based entailment. Two annotators manually annotate the relation among text-hypothesis pairs with an inter-rater agreement of 0.80. We compare the performance of the rules which gave an overall accuracy of 0.83. Further, we compare the performance of textual entailment under various conditions and the overall accuracy was 89.54%, 90.00% and 96.19% for our three datasets respectively.

1. Introduction

Argument mining (Abbas and Sawamura, 2008; Palau and Moens, 2009) deals with the extraction of argument components and structures from natural language texts. In computational argumentation, an argument can be defined as a collection of premises together (linked argument) or individually (convergent argument) which are related to a conclusion (Palau and Moens, 2009). Each premise provides a *support* in the form of logical reasoning for, or evidence in support of, the conclusion to which it is connected.

It has been suggested that, in natural language texts, this support relation can be interpreted as meaning either (a) one premise is inferred from another premise (Janier et al., 2014) or (b) one premise provides evidence that supports another premise (Park and Cardie, 2014). In either case, it is natural to interpret the relationship as a form of entailment.

In this paper, we consider a subtype of entailment, which we term as *support-based entailment*, where a support relation exists between the text and the hypothesis. We create a dataset of text-hypothesis pairs from opinions collected from a set of hotel reviews where the text provides support to the corresponding hypothesis. Human annotation of argument structures and relation among them is a complicated task which is domain-dependent and hence manually annotating huge data is costly and difficult (Matthias and Stein, 2016). To achieve this, we use a distant supervision approach by manually creating a set of rules based on meta-linguistic attributes such as *stance*, *sentiment* and *specificity*. These rules automatically label a set of sentences, which is then used to train a classifier for predicting support-based entailment.

2. Support-based entailment

The three components of the proposed method are explained below. Based on these, we manually identify a set of support-based entailment rules (SER) for predicting the support-based entailment between a text(T) and a hypothesis(H).

Opinion and Premise: We take an *opinion* to be a sentence-level statement, which might be either positive or negative in sentiment, and talks about an aspect or several aspects of a product/service. For example, *service*, *location* are aspects of hotels in the hotel domain.

We consider a *premise* as a simple atomic unit that talks about one particular aspect. Hence, any opinion that talks about several aspects can be considered as a collection of several premises that may or may not be related.

Sentiment: The positive/negative sentiment of an opinion is taken into consideration. We ignore objective opinions as it cannot be used to match the global sentiment (overall star rating). As a first step we only consider TH pairs as opinions with the same sentiment.

Stance: Previously (Rajendran et al., 2017), we explain how to classify the stance expressed by an opinion as implicit/explicit. In both, the stance (for/against) is expressed by the reviewer. But, explicit opinions have the stance explicitly expressed using (1) direct approval/disapproval or (2) words/phrases by the reviewer that have a stronger intensity of expression with respect to the topic in discussion. General cues such as *recommend*, *great*, *worst* indicate direct expressions and are useful in identifying explicit opinions. Specific cues that are related to domain-based targets can help in identifying implicit opinions. For example, *lightweight laptop* has a positive stance towards the target *laptop* whereas *the storyline of the book is lightweight* has a negative stance towards the target *book*. Also, opinions can express justification such as reasons that express stance implicitly. An example is provided in Fig. 1.

Specificity: A knowledge base (KB) is created based on the domain and the aspects present where one aspect is a sub-class of the other. Given such a KB, we describe three domain-based ontology relations between two premises that make use of the implicit/explicit nature of the opinions in which the aspects are present.

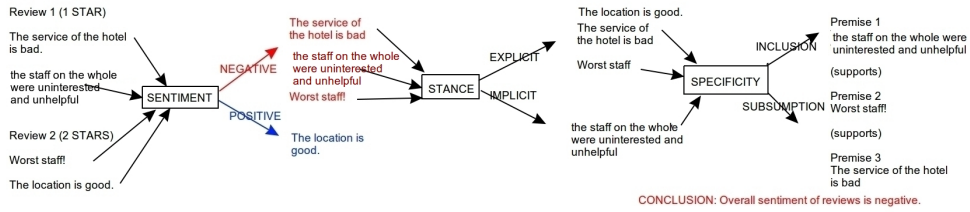

Figure 1: Opinions from two reviews are extracted and distinguished based on their local sentiment, stance, and specificity. All opinions that do not match the overall sentiment of the reviews are discarded. The rest of the opinions are then, classified as explicit or implicit and using subsumption and inclusion relation, these opinions are combined such that one supports another.

Rule	# Aspects (Text)	#Aspects (Hypothesis)	Text	Hypothesis	Relation
Rule 1	>1	>1	$a \sqsubseteq_{intrasub} b$	$c \sqsubseteq_{intrasub} d$	$b \sqsubseteq_{intersub} d$ or $b \equiv d$ and $a \sqsubseteq_{intersub} c$ or $a \equiv c$
Rule 2	>1	1	$a \sqsubseteq_{intrasub} b$	c	$b \sqsubseteq_{intersub} c$ or $b \equiv c$
Rule 3	>1	1	*a,b* and not related	c	$a \sqsubseteq_{intersub} c$ and $b \sqsubseteq_{intersub} c$
Rule 4	>1	1	*a,b* and not related	c	$a \equiv c$ or $b \equiv c$
Rule 5	1	1	a	c	$a \sqsubseteq_{intersub} c$
Rule 6	1	1	a	c	$a \equiv c$
Rule 1	1	1	a	c	$a \sqsubseteq_{inc} b$
Rule 2	1	>1	a	$b \sqsubseteq_{intrasub} c$	$a \sqsubseteq_{inc} b$
Rule 3	>1	>1	*a,b* and not related	$c \sqsubseteq_{intrasub} d$	$a \sqsubseteq_{inc} c$ and $b \sqsubseteq_{inc} d$
Rule 4	>1	1	*a,b* and not related	c	$a \sqsubseteq_{inc} c$ or $b \sqsubseteq_{inc} c$
Rule 5	1	>1	a	*b,c* and not related	$a \sqsubseteq_{inc} b$ or $a \sqsubseteq_{inc} c$
Rule 6	>1	>1	*a,b* and not related	*c,d* and not related	$a \sqsubseteq_{inc} c$ or $b \sqsubseteq_{inc} d$

Table 1: Each proposed rule for subsumption (top) and inclusion (bottom) relation is presented. The number of aspects (premises) that must be present in text and hypothesis is given. Conditions that must hold true in text, hypothesis and between them is also given. Here, we consider *a*,*b*,*c* and *d* to represent the aspects (premises) present.

Rule	Text	Hypothesis	Relation
Rule 1	and the **service** from the **staff** was extremely poor ($staff_{text} \sqsubseteq_{intrasub} service_{text}$)	it is the worst **service** i have seen in a five star **hotel** ($service_{hyp} \sqsubseteq_{intrasub} hotel_{hyp}$)	$service_{text} \sqsubseteq_{intersub} hotel_{hyp}, staff_{text} \sqsubseteq_{intersub} service_{hyp}, service_{text} \equiv service_{hyp}$
Rule 2	**location** of the **hotel** is really well placed - you're in the middle of everything ($location_{text} \sqsubseteq_{intrasub} hotel_{text}$)	overall a very good **hotel** ($hotel_{hyp}$)	$hotel_{text} \equiv hotel_{hyp}$
Rule 3	weak **service** for very high **prices** ($service_{text}, prices_{text}$)	i would not plan to stay at this **hotel** again ($hotel_{hyp}$)	$service_{text} \sqsubseteq_{intersub} hotel_{hyp}, prices_{text} \sqsubseteq_{intersub} hotel_{hyp}$
Rule 4	weak **service** for very high **prices** ($service_{text}, prices_{text}$)	however this is probably the worst **service** we have ever experienced ($service_{hyp}$)	$service_{text} \equiv service_{hyp}$
Rule 5	great **location** ($location_{text}$)	i absolutely loved this **hotel** ($hotel_{hyp}$)	$location_{text} \sqsubseteq_{intersub} hotel_{hyp}$
Rule 6	i absolutely loved this **hotel** ($hotel_{text}$)	overall a very good **hotel** ($hotel_{hyp}$)	$hotel_{text} \sqsubseteq_{intersub} hotel_{hyp}$
Rule 1	**hotel** infrastructure is in need of serious upgrading ($hotel_{text}$)	so believe me when i say do not stay at this **hotel** ($hotel_{hyp}$)	$hotel_{text} \sqsubseteq_{inc} hotel_{hyp}$
Rule 2	the **staff** that we encountered were very friendly and helpful ($staff_{text}$)	and the **service** from the valet and front desk **staff** is very good ($staff_{hyp} \sqsubseteq_{intrasub} service_{hyp}$)	$staff_{text} \sqsubseteq_{inc} staff_{hyp}$
Rule 4	to their credit the management was more responsive and very apologetic for the condition of my **room** and the rude treatment by their **staff** ($room_{text}, staff_{text}$)	dissapointed from the **room** ($room_{hyp}$)	$room_{text} \sqsubseteq_{inc} room_{hyp}$
Rule 5	the **staff** was not friendly nor helpful ($staff_{text}$)	overall its a dark dated **hotel** let down badly by the unhelpful and rude **staff** ($hotel_{text}, staff_{hyp}$)	$staff_{text} \sqsubseteq_{inc} staff_{hyp}$

Table 2: Examples for different rules satisfying subsumption (top) and inclusion (bottom) relations.

Suppose an aspect is present in a given opinion, we consider the opinion to contain a premise about that particular aspect. We thus represent each such premise as $\mathcal{P}(attr, op, stance)$ where $attr$ is the *aspect* present in an opinion *Op* which is classified as implicit/explicit and represented as *Stance*. We define the three relations below.

Def. 1 (Subsumption, $\sqsubseteq_{sub}$). Two premises present within an opinion, $\mathcal{P}(attr1, op1, exp) \sqsubseteq_{intrasub}$ (intra-subsumption) $\mathcal{P}(attr2, op1, exp)$ if *attr1* is a sub-class of *attr2*.
Two premises present in two different opinions, $\mathcal{P}(attr1, op1, exp) \sqsubseteq_{intersub}$ (inter-subsumption) $\mathcal{P}(attr2, op2, exp)$ if *attr1* is a sub-class of *attr2*.

Def. 2 (Inclusion, $\sqsubseteq_{inc}$). Two premises, one present in an implicit opinion and the other present present in an explicit opinion satisfies $\mathcal{P}(attr1, op1, imp) \sqsubseteq_{inc}$ (is-inclusive of) $\mathcal{P}(attr2, op2, imp)$ such that *attr1* and *attr2* are the same.

Def. 3 (Equivalence, $\equiv$). $\mathcal{P}(attr1, op1, exp) \equiv$ (equivalent) $\mathcal{P}(attr2, op2, exp)$ if *attr1* and *attr2* are same. $\mathcal{P}(attr1, op1, imp) \equiv$ (equivalent) $\mathcal{P}(attr2, op2, imp)$ if *attr1* and *attr2* are same.

2.1. Support-based entailment rules

Our definition of a premise states that an opinion with *n* aspects contains *n* premises. We are not interested in decomposing the opinion into different premises based on the linguistic structure but instead focus on identifying text-hypothesis (TH) pairs. Our motivation behind creating the dataset is to identify TH pairs that can help in forming argument structures from these premises using implicit and explicit opinions. A simple structure would be of the form $(implicit_1, explicit_1, explicit_2)$ with different relations as follows:

- Inclusion relation between a premise present in $implicit_1$ and a premise in $explicit_1$. Both premises are about the same aspect.
- Intra-subsumption relation between two different premises present within $explicit_1$. Same can be said for $explicit_2$.
- Inter-Subsumption/Equivalence relation between a premise in $explicit_1$ and a premise in $explicit_2$.

All these relations require two premises. For every opinion (text or hypothesis), our rules are designed to consider atmost two premises at a time and whether those two premises are related or not. Also, if an opinion contains more than one premise, then rules based on a single premise cannot be considered. For example, let us consider a text that contains 3 premises *a,b* and *c* with *a* and *b* related. For a given hypothesis, one rule will satisfy based on the related premises *a* and *b* while some other rule might satisfy based on two premises that are not related (eg. *a* and *c*). We predict the support-based entailment in a TH pair if atleast one of the rules is satisfied. This is to ensure that there are no duplicate pairs created.

Data	Rev	Exp	Imp	Sub	Inc
FA	369	264	720	Rule 1: 14	Rule 1: 271
				Rule 2: 138	Rule 2: 25
				Rule 3: 27	Rule 3: 6
				Rule 4: 218	Rule 4: 619
				Rule 5: 193	Rule 5: 147
				Rule 6: 218	Rule 6: 344
SA	707	1001	4359	Rule 1: 92	Rule 1: 1790
				Rule 2: 566	Rule 2: 137
				Rule 3: 82	Rule 3: 55
				Rule 4: 344	Rule 4: 3418
				Rule 5: 842	Rule 5: 933
				Rule 6: 1834	Rule 6: 1799
UA	3271	564	5933	Rule 1: 34	Rule 1: 3708
				Rule 2: 467	Rule 2: 148
				Rule 3: 55	Rule 3: 33
				Rule 4: 119	Rule 4: 4726
				Rule 5: 428	Rule 5: 2189
				Rule 6: 1354	Rule 6: 3053

Table 3: In each dataset: total number of reviews (Rev) present, total number of explicit opinions (Exp) and implicit opinions (Imp) found and total number of TH pairs satisfying each rule in SER based on subsumption (Sub) and inclusive (Inc) relation is present.

If a text/hypothesis can contain a single premise or atmost two premises, then 9 different combinations based on whether inter-subsumption is present in the text/hypothesis or not. Based on our definition of *support-based entailment*, a specific premise supports a more generalised premise. Thus, we ignore rules based on subsumption relation that look into hypothesis containing non-related premises. So, we have an overall of 6 different combinations. Also, implicit opinions (text) cannot have any inter-subsumption relation and hence 3 combinations are ignored. Thus, we have an overall of six different rules based on inclusion relation. These rules are present in Table 1.
Given two explicit opinions of same sentiment, we apply the rules based on subsumption relation. Fistly, we check for intra-subsumption related premises within each text and hypothesis and apply the corresponding rules. If not, rules based on unrelated and single premises are applied. Given an implicit opinion and an explicit opinion of same sentiment, we apply the rules based on inclusion relation. Single premises within the text and hypothesis are checked first and the corresponding rules are applied. Otherwise, hypothesis with related premises is considered and the rule is applied. Then, text and hypothesis with unrelated premises are considered and the rules are applied accordingly.

3. SSS (Sentiment-Stance-Specificity) Dataset

We use an existing hotel reviews corpus, ArguAna (Wachsmuth et al., 2014b) to create our datasets. The data for each hotel contains a balanced set of reviews based on the overall star rating for that hotel. Each review contains manually annotated local sentiment of the statements (pos, neg or obj), aspects present and the overall star rating.

First, we create a knowledge base using a list of aspects extracted from the ArguAna corpus. For example, *(Location $\sqsubseteq_{sub}$ Hotel), (Service $\sqsubseteq_{sub}$ Hotel), (Cleanliness $\sqsubseteq_{sub}$ Hotel), (Staff $\sqsubseteq_{sub}$ Service), (Restaurant service $\sqsubseteq_{sub}$ Service) etc.*

We used the manually annotated 1288 implicit/explicit opinions dataset created in (Rajendran et al., 2017) which was annotated by two annotators with an inter-rater agreement of Cohen's Kappa = 0.70. Finally, three different dataset were created for our experiment using the proposed rules (few examples in Table 2):

1. **Fully annotated (FA)** Reviews from 15 different hotels balanced based on overall star ratings. Local sentiment of statements, aspects present and implicit/explicit classification are manually annotated.

2. **Semi-annotated (SA)** Reviews from 32 different hotels balanced based on overall star ratings. Extracted opinions are automatically classfied as implicit/explicit using an SVM-based classifier with features mentioned in (Rajendran et al., 2017).

3. **Unannotated (UA)** Unannotated and unbalanced set of reviews not present in ArguAna extracted from 26 different hotels. Local sentiment of each statement is automatically classfied as *pos*,*neg* or *obj* using the SVM-based classfier described in (Wachsmuth et al., 2014a). Aspects manually annotated in the ArguAna corpus are used to identify aspects in this dataset. The opinions are automatically classfied as implicit/explicit as mentioned in previous dataset.

4. Performance of SER

In each of the above datasets, we predict the support-based entailment relation using the SER and present the total numbe of predicted cases in Table 3. We extracted 160 TH pairs based on the SER as well as those that do not satisfy them. Two annotators were manually asked to annotate whether the pairs satisfy support-based entailment or not. No information about the rules were provided. The inter-rater agreement was calculated using Cohen's Kappa as 0.80. To test the performance of the SER, we took the intersection of the two annotations as the groundtruth data and the accuracy of the SER prediction was 0.83. We also considered the union of the two annotations as the groundtruth data which gave the accuracy of the SER prediction as 0.93.

4.1. Performance of textual entailment

We use the Excitement Open Platform (EOP) (Magnini et al., 2014) to automatically predict textual entailment in support-based entailment relation. The EOP tool takes a text and a hypothesis as input and predicts whether text (T) entails the hypothesis (H) or not. We use the TH pairs that are predicted as support-based entailment using the 12 different SER (Table 1). The MaxEntClassificationEDA (Magnini et al., 2014) which is based on the maximum entropy classifier gave the best performance with the RTE-3 (Giampiccolo et al., 2007)s dataset and overall accuracy of 89.54 % on the *FA* dataset and hence we use this classifier and the training data for our experiments.

Experiment	FA	SA	UA
SER	89.54	90.00	96.19
Non-SER	76.18	72.69	88.01
Subsumption based SER	81.63	75.82	92.11
Subsumption based Non-SER	73.91	67.93	86.21
Inclusion based SER	95.83	96.49	97.68
Inclusion based NON-SER	76.87	73.84	88.31
Implicit-Explicit Entailment	75.94	71.03	87.89
Subsumption			
-Rule 1	100.0	83.69	100.0
-Rule 2	86.95	92.40	96.14
-Rule 3	44.44	52.43	80.0
-Rule 4	89.44	93.89	99.15
-Rule 5	62.69	46.67	83.64
-Rule 6	86.69	81.35	92.17
Inclusion			
-Rule 1	92.61	93.74	94.76
-Rule 2	96.0	95.62	96.62
-Rule 3	100.0	94.59	100.0
-Rule 4	97.25	98.50	98.47
-Rule 5	89.79	92.60	95.56
-Rule 6	95.63	97.72	98.59
Random sentiment (SER)	45.62	45.31	47.98
Random sentiment (Non-SER)	38.64	36.37	44.02

Table 4: Experiment is run on each dataset by (a) SER - TH pairs satisfying either of the six subsumption or six inclusion rules (b) Non-SER - TH pairs that do not satisfy any of the 12 rules. (c) Subsumption and Inclusion - TH pairs satisfying each individual rule and (d) Random sentiment - Assigning randomly sentiment of opinions present in TH pairs of SER and Non-SER. Accuracy is reported.

We evaluate the performance of automatically predicting entailment by conducting different set of experiments on the three different datasets and the accuracy of correct prediction in each of these experiments is listed in Table 4. As observed from Table 4, our method is effective for support-based entailment prediction in all three datasets as the overall accuracy of SER outperforms that of Non-SER. From the results of the individual rules, it is evident that textual entailment does not depend on the domain knowledge base and does not consider specificity as a property for prediction. We also experimented by randomly assigning incorrect sentiment (random sentiment baseline) and as expected the accuracy was lowered in comparison with SER.

5. Conclusion

We present three datasets of TH pairs based on a subtype of entailment, which we term as support-based entailment that predicts the support relation between a specific premise and a generalised premise using sentiment, stance and specificity. A distant supervision approach is carried out by using a set of proposed rules based on three components – *sentiment*, *stance* and *specificity*. The performance of these rules against manually annotated 160 TH pairs is measured by the accuracy as 0.83. Experiments on the three datasets for textual entailment task shows that the rules are able to predict the entailment relation but existing textual entailment method is not able to capture support-based entailment. We believe that our datasets will be useful to expedite research in argument mining.

6. Bibliographical References

Abbas, S. and Sawamura, H. (2008). A first step towards argument mining and its use in arguing agents and its. In *KES*, pages 149–157.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The third pascal recognizing textual entailment challenge. In *ACL-PASCAL*, pages 1–9.

Janier, M., Lawrence, J., and Reed, C. (2014). Ova+: an argument analysis interface. In *COMMA*, pages 463–464.

Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Pado, S., Stern, A., and Levy, O. (2014). The excitement open platform for textual inferences. In *ACL*, pages 43–48.

Matthias, K. A.-K. H. W. and Stein, H. J. K. B. (2016). Cross-domain mining of argumentative text through distant supervision. In *NAACL-HLT*, pages 1395–1404.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *ICAIL*, pages 98–107.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *ACL*, pages 29–38.

Rajendran, P., Bollegala, D., and Parsons, S. (2017). Identifying argument based relation properties in opinions. In *PACLING*, page to appear.

Wachsmuth, H., Trenkmann, M., Stein, B., and Engels, G. (2014a). Modeling review argumentation for robust sentiment analysis. In *COLING*, pages 553–564.

Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., and Palakarska, T. (2014b). A review corpus for argumentation analysis. In *CICLing*, pages 115–127.