SRL4ORL: Semantic Role Labelling for Opinion Role Labelling

Ana Marasović
Department of Computational Linguistics
Heidelberg University

AIPHES
ADAPTIVE PREPARATION OF INFORMATION FROM HETEROGENEOUS SOURCES

RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Heidelberger Institut für Theoretische Studien

HITS

TECHNISCHE UNIVERSITÄT DARMSTADT

Fine-grained opinion analysis (FGOA)

FGOA aims to

Fine-grained opinion analysis (FGOA)

FGOA aims to

- detect explicit **opinion expressions** (O)
- measure their *intensity* (e.g. strong)

John likes that she enjoys being at the Enderly Park.

O_1 O_2

Fine-grained opinion analysis (FGOA)

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

FGOA aims to

- detect explicit **opinion expressions** (O)
- measure their *intensity* (e.g. strong)
- identify their **targets** (T), entities or propositions at which sentiment is directed

John likes that $\boxed{\text{she enjoys being at the Enderly Park}}$.

O_1 $\mathbf{T_1}$

John likes that she enjoys $\boxed{\text{being at the Enderly Park}}$.

O_2 $\mathbf{T_2}$

Fine-grained opinion analysis (FGOA)

FGOA aims to

- detect explicit **opinion expressions** (O)
- measure their *intensity* (e.g. strong)
- identify their **targets** (T), entities or propositions at which sentiment is directed
- identify their **holders** (H), entities that express an opinion

John likes that $\boxed{\text{she enjoys being at the Enderly Park}}$.

H_1 O_1 T_1

John likes that she enjoys $\boxed{\text{being at the Enderly Park}}$.

H_2 O_2 T_2

Fine-grained opinion analysis (FGOA)

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

FGOA aims to

- detect explicit **opinion expressions** (O)
- measure their *intensity* (e.g. strong)
- identify their **targets** (T), entities or propositions at which sentiment is directed
- identify their **holders** (H), entities that express an opinion
- classify *target-dependent* sentiment they express toward their targets

positive

John likes that $\boxed{\text{she enjoys being at the Enderly Park}}$.

$\mathbf{H_1}$ O_1 T_1

positive

John likes that she enjoys $\boxed{\text{being at the Enderly Park}}$.

$\mathbf{H_2}$ O_2 T_2

Outline

✓Fine grained-opinion analysis

What are the aims of fine-grained opinion analysis?

- Related work

How is FGOA approached?

What kind of data is commonly used?

- Semantic Role Labelling (SRL) for Opinion Role Labelling (ORL)

Could we adapt SRL models for ORL?

How well can we predict opinion expressions only?

Could we exploit SRL data?

Preliminary results

- Future directions and discussion

Approaches to FGOA

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

- span-based annotated MPQA corpus (Wiebe et al. **2005**)
  ⇒ **sequence tagging** models with the BIO encoding scheme

John likes that she enjoys being at the Enderly Park .

B-H	B-O	O	B-T	I-T	I-T	I-T	I-T	I-T	I-T
O	O	O	B-H	B-O	B-T	I-T	I-T	I-T	I-T
B-H	B-O	O	B-H	B-O	B-T	I-T	I-T	I-T	I-T

standard methods, CRF and recurrent nets, produce just one of these sequences

- **pipeline models**
  - first label opinion expressions and then, given an opinion, label its holders and targets (*opinion roles*) (Kim and Hovy, 2006; Kobayashi et al., 2007)
  - overlapping entities are handled
  - target-dependent sentiment classification not done

Joint inference models for FGOA

Joint inference: labelling of *opinion entities* (expressions, holders, targets) and classification of relations between them into

- *is-about*: from a target to its opinion expressions
- *is-from*: from an opinion to its holder

$\Big\}$ *opinion relations*

- **Yang and Cardie (2013)** (CRF + ILP)

$$\arg\max_{x,u,v} \lambda \sum_{i \in \mathcal{S}} \sum_{z} f_{iz} x_{iz} \; + (1-\lambda) \sum_{k} \sum_{i \in \mathcal{O}} \left(\sum_{j \in \mathcal{A}_k} r_{ij} u_{ij} + r_{i\emptyset} v_{ik} \right)$$

  - **state-of-the-art**
  - discard entities that contain other entities
  - without target-dep. sentiment classification

The most recent neural models for FGOA

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

Katiyar and Cardie (2016) (LSTM + RLL)

WLL = word-level log likelihood ⇒ standard LSTM

SLL = sentence-level log-likelihood ⇒ the best sequence of opinion entity labels

RLL = relational-level log likelihood ⇒ the best sequence of opinion entity labels and

the best sequences of relational distances

is-from is-from is-about

John likes that she enjoys being at the Enderly Park .

B-H	B-O	O	B-H	B-O	B-T	I-T	I-T	I-T	I-T
0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	2	3	4	5

❗ CRF + ILP outperforms LSTM + RLL for extraction of opinion roles

❗ CRF - ILP outperforms LSTM + SLL baseline for extraction of opinion roles

- discard entities that contain other entities, without target-dep. sentiment class.

Where are we 12 years after?

TECHNISCHE UNIVERSITÄT DARMSTADT

AIPHES

- neural models lag behind feature-based models
- (constantly) left as future work
  - handling opinion entities that contain other opinion entities
  - target-dependent sentiment classification

is-from is-about is-from is-about

John	likes	that	she	enjoys	being	at	the	Enderly	Park	.
B-H	B-O	O	B-H / B-T	B-O / I-T	B-T / I-T	I-T	I-T	I-T	I-T	
0	0	0	0	0	0	0	0	0	0	
0	1	0	0 / 2	1 / 3	1 / 4	2 / 5	3 / 6	4 / 7	5 / 8	

❗ CRF+ILP & LSTM+RLL achieve ~54% F1 score for classification of *is-about* rel.

⇒ **the models we have at the moment are not close to answering:**

Who expressed what kind of sentiment towards what?

How can we improve?

Can we build simpler models? Using more data?

1. SRL has substantially more data

	train	dev	test-WSJ	test-Brown	test
CoNLL'05	90750	3248	5269	804	6073
MPQA	3458	1224	-	-	313

2. SRL is similar in nature to ORL

<table>
<tr><th></th><th>John</th><th>likes</th><th>that</th><th>she</th><th>enjoys</th><th>being</th><th>at</th><th>the</th><th>Enderly</th><th>Park</th><th>.</th></tr>
<tr><td>like.01</td><td>A0</td><td></td><td colspan="8">A1</td><td></td></tr>
<tr><td>enjoy.01</td><td colspan="3"></td><td>A0</td><td></td><td colspan="5">A1</td><td></td></tr>
</table>

The output of Semantic Role Labeling demo. Looks familiar?

SRL and ORL: differences

- **related work**
  - mapping from semantic to opinion roles: Kim and Hovy (2006)
  - SRL frame as a feature: Choi et al. (2006)
  - ORL poses challenges beyond SRL: Wiegand and Ruppenhofer (2015)

Peter$_{agent}$ criticized Mary$_{patient}$. ⇒ (criticize, Peter, holder) & (criticize, Mary, target)
Peter$_{agent}$ disappoints Mary$_{patient}$. ⇒ (disappoint, Peter, target) & (disappoint, Mary, holder)

- **our perspective:** how could SRL resources (models, data) be exploited for ORL?
- **first step:** adapt state-of-the-art SRL model (Zhou and Xu, 2015) for ORL
- **challenges:**
  - ! SRL models (usually) presuppose that a **predicate is given**
  - we can not use sentiment lexical resources to extract opinion expressions:
    - *holding himself accountable, demonstrate his concern*
    - *asked, said…*

ZX-SRL model (Zhou and Xu, 2015)

John	likes	that	she	enjoys	being	at	the	Enderly	Park	.
B-A0	B-P	B-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	I-A1	
O	O	O	B-A0	B-P	B-A1	I-A1	I-A1	I-A1	I-A1	
~~B-A0~~	~~B-P~~	~~O~~	~~B-A0~~	~~B-P~~	~~B-A1~~	~~I-A1~~	~~I-A1~~	~~I-A1~~	~~I-A1~~	

process a sentences as many times as there are predicates in it

word vector = *concatenate*(token emb.,
predicate emb.,
predicate-context emb.,
isInContext)

input to bi-LSTM

predicate-context embedding = average of embeddings of words in a surrounding window of the predicate (context)

isInContext = 1 if the token is in the context, 0 otherwise

+ 1-4 layers of bi-LSTM or bi-GRU
+ CRF layer

ZX-ORL: ZX-SRL for ORL

John likes that she enjoys being at the Enderly Park .

B-H B-O O B-T I-T I-T I-T I-T I-T I-T
O O O B-H B-O B-T I-T I-T I-T I-T
~~B-H B-O O B-H B-O B-T I-T I-T I-T I-T~~

process a sentences as many times as there are opinions in it

word vector = *concatenate*(token emb.,
opinion expression emb.,
opinion expression-context emb.,
isInContext)

input to bi-LSTM

opinion expression-context embedding = average of embeddings of words in a surrounding window of the opinion expression(context)

isInContext = 1 if the token is in the context, 0 otherwise

+ 1-4 layers of bi-LSTM or bi-GRU
+ CRF layer (with a new set of labels)

Oversampling simplifies target-dependent sentiment classification

AIPHES

John likes that she enjoys being at the Enderly Park .

John	likes	that	she	enjoys	being	at	the	Enderly	Park
B-H	B-O	O	B-T	I-T	I-T	I-T	I-T	I-T	I-T
O	O	O	B-H	B-O	B-T	I-T	I-T	I-T	I-T

process a sentences as many times as there are opinions in it

1)

John likes that she enjoys being at the Enderly Park .

H_1 O_1 T_1

John likes that she enjoys being at the Enderly Park .

H_2 O_2 T_2

Oversampling simplifies target-dependent sentiment classification

John	likes	that	she	enjoys	being	at	the	Enderly	Park	.
B-H	B-O	O	B-T	I-T	I-T	I-T	I-T	I-T	I-T	
O	O	O	B-H	B-O	B-T	I-T	I-T	I-T	I-T	

process a sentences as many times as there are opinions in it

1)

John likes that she enjoys being at the Enderly Park .

H_1 O_1 T_1

John likes that she enjoys being at the Enderly Park .

H_2 O_2 T_2

2)

positive

John likes that she enjoys being at the Enderly Park .

H_1 O_1 T_1

positive

John likes that she enjoys being at the Enderly Park .

Exp. 1: ZX-ORL with gold opinion expressions

- **dataset:** MPQA 2.0.
- **model:** ZX-ORL with gold opinion expressions
- **validation technique:** 10-fold CV
- **evaluation metrics:**
  - *binary overlap recall*: how many gold entities have an overlapping predicted entity
  - *binary overlap precision*: how many predicted entities have an overlapping gold entity
  - *proportional overlap recall*: measures proportion of the overlap between a gold entity and an overlapping predicted entity
  - *proportional overlap precision*: measures proportion of the overlap between a predicted entity and an overlapping gold entity
  - f-score = $\frac{2 * precision * recall}{precision + recall}$
- **manually chosen HPs**

Exp 1: ZX-ORL with gold opinion expressions -- results

	10-fold CV binary overlap f-score			10-fold CV proportional overlap f-score		
	expression	holder	target	expression	holder	target
ZX-ORL	98.30	73.40	74.08	97.41	71.06	68.84
CRF	71.17	59.21	59.19	61.67	57.86	53.23
LSTM+SLL	68.37	62.35	59.65	63.60	60.40	52.01
CRF+ILP	74.11	67.22	65.40	70.22	65.68	58.72
LSTM+RLL	71.11	64.71	64.84	65.56	62.18	55.81

- **direct comparison is not possible, our model is given gold opinion expressions!**
- ZX-ORL results serve as an upper bound for a *future* pipeline model
  - opinion extraction ⇒ role extraction ⇒ target-dependent sentiment classification

Exp 2.: Predicting opinion expressions only

- **model variants:**
  1. LSTM
  2. GRU
  3. LSTM + CRF
  4. GRU + CRF
  - shallow (1 layer) vs. deep (3 layers)

- **HPs tuned** with **Tree-structured Parzen Estimators** (Bergstra et al., 2011)
  - tuned with dev prop. f-score in 50 trails for each out of 8 architectures separately
  - the size of the LSTM/GRU hidden state, value for clipping gradients, word frequency threshold, l_2-regularization coefficient, keep input probability, keep output probability

Exp 2: Predicting opinion expressions only -- Results

TECHNISCHE UNIVERSITÄT DARMSTADT

		dev		test (only 1. fold)		
		prop. f-score	binary f-score	prop. f-score	binary f-score	# params
3 layers	LSTM	57.96	70.60	61.30	73.22	113836
	GRU	61.82	70.87	62.59	70.45	615500
	LSTM + CRF	67.19	74.48	67.64	73.23	653692
	GRU + CRF	**66.85**	**74.25**	**68.98**	**74.27**	**1390552**
1 layer	LSTM	60.46	69.94	61.69	69.41	204796
	GRU	60.52	70.73	61.65	70.16	175256
	LSTM + CRF	65.89	74.42	67.04	74.09	803220
	GRU + CRF	**67.34**	**73.36**	**69.35**	**74.12**	**458892**

- the shallow GRU + CRF achieves as good performance as the 3-layer GRU + CRF

Exp 2: Predicting opinion expressions only -- 10-fold CV results

	10-fold CV test	
	prop. f-score	binary f-score
GRU + CRF (1 layer)	66.20	73.43
Irsoy & Cardie (2014)	66.01	71.72

- predicting **SRL predicates:**
  - bi-LSTM model achieved an F1 score of 91.43% on marking words as predicates (or not) (Swayamdipta et al., 2016)
- **next step**: feed predicted opinion expressions to ZX-ORL (no results yet, sorry)

Transfer learning from SRL

How about transfer learning from SRL?

- so far: we used the model for SRL (ZX-SRL), but did not use SRL data
- **transfer learning**: pre-train ZX-SRL and **fine-tune it for ORL**
  - tuning the new last layer that outputs ORL labels (CRF layer)
  - tuning the full architecture (without the SRL-CRF layer, with the ORL-CRF layer)

Exp 3: ZX-SRL + fine-tuning

TECHNISCHE UNIVERSITÄT DARMSTADT

	5-fold CV binary overlap f-score			5-fold CV proportional overlap f-score		
	expression	holder	target	expression	holder	target
ZX-ORL	98.36	72.84	74.22	96.47	70.47	68.79
ZX-SRL + FT (full)	98.17	63.93	68.80	97.49	61.38	64.19

Exp 3: ZX-SRL + fine-tuning

	5-fold CV binary overlap f-score			5-fold CV prop. overlap f-score		
	expression	holder	target	expression	holder	target
ZX-ORL	98.36	72.84	74.22	96.47	70.47	68.79
ZX-SRL + FT (full)	98.17	63.93	68.80	97.49	61.38	64.19
ZX-SRL + FT CRF	39.53	13.75	13.38	37.34	12.13	13.11

- the new CRF layer is randomly initialized ⇒ fine-tuning is very sensitive to the choice of the learning rate
- after the first epoch
  - **with fine-tuning:** 77.35% (opinion expression), 18.59% (holder) and 56.89% (target) binary f-score
  - **without fine-tuning:** 0%, 0% and 1.17% binary f-score
  - **the pre-trained weights are relatively good, but get distorted too quickly and too much**

Future directions

- **apply more sophisticated fine-tuning techniques**
  - impact of learning rate on fine-tuning: global and local
  - layer-wise transfer
  - restrict ZX-SRL to predict only relevant roles
- **deeper analysis** to investigate the real impact of SRL for ORL
  - visualization
  - perturbation analysis
- **full pipeline model**
  - opinion extraction ⇒ role extraction ⇒ target-dependent sentiment classification
  - consider end-to-end learning

Thank you for your attention!

References

- James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NIPS). Granada, Spain.
- Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint Extraction of Entities and Relations for Opinion Recognition. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, USA.
- Ozan Irsoy and Claire Cardie. 2014. Opinion Mining with Deep Recurrent Neural Networks. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar.
- Arzoo Katiyar and Claire Cardie. 2016. Investigating lstms for joint extraction of opinion entities and relations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) . Berlin, Germany.

References

- Soo-Min Kim and Eduard Hovy. 2006. Extracting Opinions, Opinion Holders, And Topics Expressed In Online News Media Text. In Proceedings of the Workshop on Sentiment and Subjectivity in Text (SST). Stroudsburg, USA.

- Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL). Prague, Czech Republic.

- Swabha Swayamdipta, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Greedy, Joint Syntactic-Semantic Parsing with Stack LSTMs. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL). Berlin, Germany. 2016.

- Michael Wiegand and Josef Ruppenhofer. 2015. Opinion Holder and Target Extraction based on the Induction of Verbal Categories. Proceedings of the 19th Conference on Computational Language Learning (ACL). Beijing, China.

References

- Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia, Bulgaria.

MPQA corpus

- span-based annotated MPQA corpus (Wiebe et al. **2005**)

⇒ **sequence tagging** models with the BIO encoding scheme

"If the prosecution's key witness **is reluctant** to testify, how will they proceed?" Mr. Tsvangirai asked. "There is no case to answer."

- explicit opinion annotations allow annotating an implicit opinion as explicit with an "implicit" arg.
- **but**, annotators were not consistent with marking the "implicit" argument

MPQA corpus

- span-based annotated MPQA corpus (Wiebe et al. **2005**)

⇒ **sequence tagging** models with the BIO encoding scheme

Asked whether […], Tbeishat **said** that signing the protocol **will have a "positive" outcome […]**.

He **explained** that both the US and Jordan **have different issues to deal with on a national level**, including environmental issues.

- "said", "explained" are not sentiment-barring words
- **attitude** annotations: "e.g., positive sentiments, negative sentiments, agreements, etc., being expressed overall by the private states represented by the direct subjective."

Ideas for target-dependent sentiment classification

- concatenate target word embeddings ⇒ feed-forward layer of hidden units
- average target word embeddings ⇒ feed-forward layer of hidden units
- process a target sequence with some recurrent architecture or CNN
- process a sentence with some recurrent architecture and concatenate the target embedding with every token embedding