Aspect-based sentiment analysis of conference review forms with LD-enabled review criteria

Sára Juranková (VSE, Prague, Czech Republic)
Vojtěch Svátek (VSE, Prague, Czech Republic)
Chiara Ghidini (FBK, Italy)

Supported by IGA VSE 56/2021 ("Knowledge Engineering of PhD Stories on the Semantic Web") and by Nexus Linguarum (COST Action CA18209)

Overview

Research goals

- **Tentative application of** *sentiment analysis* **on conference paper reviews**
- Exploration of the connection of review forms to *linked data*
- Exploration of the social/business *context* of review forms' content

Peer review of conference papers and its relation to LD

- Importance of peer review for multiple kinds of stakeholders:
  - feedback to *authors*' research; support for *Chairs*' acceptance/rejection decision; improved *readers*' experience
- Peer reviews are mostly closed data nowadays:
  - probable reason why application of SA on reviews is scarce
  - recent push towards transparency and open data is likely to change this... thus *LOD technology* is relevant
- Mapping of reviews to a *shared set of criteria* expressed as LOD can:
  - help provide a rapid *overview* of the paper's quality
  - lead to easier *comparison of different reviews* of the same paper
  - allow for an easier comparison of reviews across *different conferences*

Structure of conference reviews vs. aspects and polarity

Different conferences have different structures of their web-based review forms:

- Usually a fixed choice for reviewer's confidence and overall evaluation
- The partial criteria ($\Rightarrow$ *aspects*) can be reflected through:
  - fixed-choice items (e.g., *5-very good*, *4-good*, *3-fair*, ...)
  - dedicated text fields
  - an overall text field; the choice of criteria covered then much depends on the reviewer
- Sometimes the comments can be separated by their *polarity* rather than by criteria (e.g. *reasons to accept/reject*)

Micro-study in the semantic technology domain

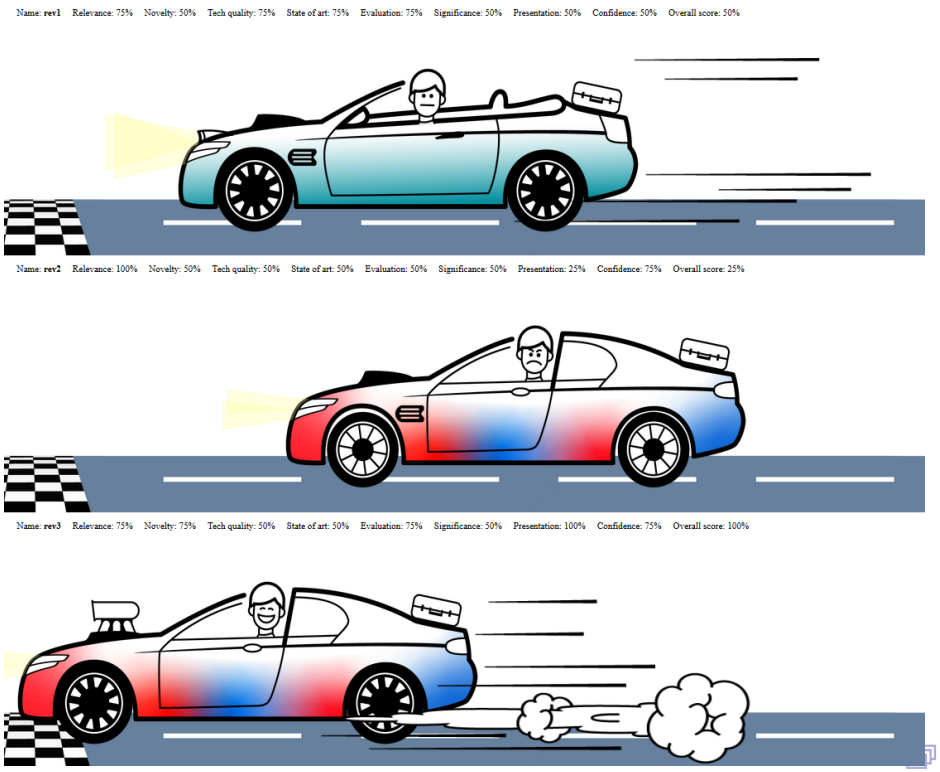
- Review forms of conferences from the semantic technology field (ECAI, EKAW, ESWC, FOIS, IJCAI, ISWC, K-CAP, KR and SEMANTiCS)
- Semantic analysis of their field labels and reviewer guidelines led to 7 partial review metrics:
  - Relevance, novelty, technical quality, state of the art, evaluation, significance, presentation
  - + two global metrics – Confidence and Overall score – present in all forms
- Wide variety of review structuring styles, while the *sets of review criteria* are semantically very similar across different conferences

Linked data infrastructure for review result sharing

- Research resulted in a prototype ontology that supports the publishing of metrics and their relationship to review forms
  - The ontology contains the classes ReviewMetrics, ReviewForm, ReviewFormField and F2M_Maping (for the field-to-metric mapping) + the connecting properties.
  - Online at `http://kizi.vse.cz/pictoreview/ontology/`
  - Proposed metrics online at `http://kizi.vse.cz/pictoreview/metrics/`
- A prototype tool suite developed that allows the user to create a mapping from the custom set of review form fields of a particular event to the proposed set of generic metric
  - Online `https://pictoreview.vse.cz/`

End application prototype: review set visualization

- Presented as demo at ISWC'20
- Multidimensional glyph metaphor approach
- Visual variables = review criteria; e.g., engine size = originality, headlamp power = SotA coverage, ...



Review forms as domain-specific material for SA

- Domain-specific features of the setting:
  - *Annotated corpora* scarce
  - *Domain-specific lexica* unavailable
  - Relatively *curated* language
  - *Explainability* desirable (if results to be used in the paper acceptance decision making)
  - Relatively free of *emotions*
  - Distinct *partial criteria* to be (nearly) mandatorily covered in the text
- Leads to the preliminary choice of *aspect-based* SA using a *newly built* domain-specific *lexicon*

Review forms as domain-specific material for SA

- Domain-specific features of the setting:
  - *Annotated corpora* scarce →**lexicon-based**
  - *Domain-specific lexica* unavailable→**build a new lexicon**
  - Relatively *curated* language→**lexicon-based**
  - *Explainability* desirable (if results to be used in the paper acceptance decision making)→**lexicon-based**
  - Relatively free of *emotions*→**build a new lexicon**
  - Distinct *partial criteria* to be (nearly) mandatorily covered in the text→**aspect-based**
- Leads to the preliminary choice of *aspect-based* SA using a *newly built* domain-specific *lexicon*

Prior research on conference paper review analysis

- Bucur et al (K-CAP, 2019):
  - Collected a dataset of eleven reviews, each of which was manually annotated by their respective authors
  - Applied 18 different lexicon-based sentiment analysis tools to compare the results and found that the best performing tool was the SOCAL method, with a max. accuracy of 72.8%
  - SA performed on the level of comments; sentiment was *not* determined at the aspect level
- Follow-up research by same authors focused on creating a unified model for representation of *publications* and their assessments in the format of linked data in order to give more context to reviews

Our data source

- Samples of data from five conferences: EKAW 2018, ESWC 2018, 163 ESWC 2019, ISWC 2017 and ISWC 2018
- ESWC 2019: linked open data, already published as anonymized; all reviews available this way
- Others: closed data in review systems; review authors gave their explicit consent; data accessed via Chair or Senior PC access, anonymized; 284 reviews extracted this way

Development of a tool for aspect-based sentiment analysis of conference reviews

Development of a tool for SA of conference reviews done in 3 major steps:

1. *Aspect expressions* extraction
2. Creation of a custom (domain-specific) *sentiment lexicon*
3. Implementation of an aspect-based *sentiment analysis algorithm*

Aspect expressions extraction

- Creation of a lexicon of aspect/criterion expression using two approaches:
  - Taxonomy extraction based on identifying *frequent noun phrases* similar to terms in a *manually created seed taxonomy*
  - Extraction of frequent words in reviews which are already divided by headers into sections for the *respective criterion*

Creation of sentiment lexicon

- Experiments with existing sentiment lexica showed a need for a creation of a domain-specific lexicon
- Created using a Naïve Bayes classifier on a set of sentences manually labeled with their polarities, chosing the features/terms with the highest contribution to the classification
  - 1000 sentences from the ESWC 2019 corpus, manually labeled by two annotators

Aspect-based sentiment analysis algorithm

- For each aspect expression in a sentence, its sentiment score is calculated using the polarity of the opinion words and their distance in the sentence from the feature expression
  - Handles negations, but-words and other modifiers
  - Inspired by the holistic lexicon-based approach (Ding, 2008) and by the sentimentr method, `https://github.com/trinker/sentimentr`
- Additional fallback rules: adjectives as aspect expressions, intra sentence rules, neutral sentiment
- Final score for a criterion is a sum function over the sentiment polarities of its aspect expressions found in the text

Evaluation using reviews with numerical scores

- 20 reviews from ISWC 2018; 6 criteria
- Numerical scores of each criterion estimated by the algorithm were compared to the scores given to these criteria by the reviewers $\Rightarrow$ MAE 0.99 on a scale [-2, 2]
- Often, when a numerical score was given for a criterion, it was not further expanded on in the review's text

Evaluation using annotated review comments

- 136 comments from 15 reviews (3 different conferences)
- Evaluation of the criterion identification
  - Precision 57,38%, recall 53.44%
  - Very diverse results between different criteria

Criterion	Precision	Recall
Relevance	50	35.3
Novelty	63.6	25.9
Technical quality	36.4	14.3
State of the art	33.3	31.3
Evaluation	57.9	52.4
Presentation	74.3	49.1

- Evaluation of polarity detection:
  - Over 75.7% of comments with correctly identified criterion were also correctly classified

Discussion of results

- The error rate is quite high for *criterion identification*, however even the annotators had a substantial level of disagreement (initially disagreeing in over a third of their annotations)
- The result for the sentiment analysis represents an improvement over the study by Bucur et al. (not aspect based, which had possibly made determining the sentiment easier)
- Use of linked data for the moment restricted to the 'factual' semantics of the criteria; not yet clear what the major added value of *linguistic* LD could be in this task
  - ...regarding the flexibility of data structures...?
  - ...regarding the interoperability of applications...?