What is SemEval evaluating? A Systematic Analysis of Evaluation Campaigns in NLP

Oskar Wysocki[1,2], Malina Florea[1], Dónal Landers[2], and André Freitas[1,2,3]

[1]Department of Computer Science, The University of Manchester
[2]digital Experimental Cancer Medicine Team, Cancer Biomarker Centre, CRUK Manchester Institute, University of Manchester
[3]Idiap Research Institute

Abstract

SemEval is the primary venue in the NLP community for the proposal of new challenges and for the systematic empirical evaluation of NLP systems. This paper provides a systematic quantitative analysis of SemEval aiming to evidence the patterns of the contributions behind SemEval. By understanding the distribution of task types, metrics, architectures, participation and citations over time we aim to answer the question on what is being evaluated by SemEval.

1 Introduction

A large portion of the empirical methods in Natural Language Processing (NLP) are defined over canonical text interpretation tasks such as Named Entity Recognition (NER), Semantic Role Labeling (SRL), Sentiment Analysis (SA), among others. The systematic creation of benchmarks and the comparative performance analysis of resources, representations and algorithms is instrumental for moving the boundaries of natural language interpretation. SemEval (May et al., 2019; Apidianaki et al., 2018; Bethard et al., 2017, 2016; Nakov et al., 2015; Nakov and Zesch, 2014; Manandhar and Yuret, 2013; Agirre et al., 2012) is the primary venue in the NLP community for the organisation of shared NLP tasks and challenges. SemEval is organised as an annual workshop co-located with the main NLP conferences and has attracted a large and growing audience of task organisers and participants.

Despite its recognition as a major driver in the creation of gold-standards and evaluation campaigns, there is no existing meta-analysis which interprets the overall contribution of SemEval as a collective effort. This paper aims to address this gap by performing a systematic descriptive quantitative analysis of 96 tasks encompassing the SemEval campaigns between 2012-2019. This study targets understanding the evolution of SemEval over this period, describing the core patterns with regard to task popularity, impact, task format (inputs, outputs), techniques, target languages and evaluation metrics.

This paper is organised as follows: section 2 describes related work; 3 describes the methodology; 4 defines the underlying task macro-categories; 5 and 6 presents the number of tasks and popularity in 2012-2019; 7 discusses SemEval impact in terms of citations; 8 shows targeted languages; then, sections 9, 10, 11 analyse input, output and evaluation metrics; 11 focuses on sentiment analysis architectures and representations; this is followed by a Discussion section; we close the paper with Recommendations and Conclusions.

2 Related work

Each SemEval task is described by an *anthology*, which contains: a summary of previous editions or similar tasks, references to previous works, detailed task description, evaluation methods, available resources, overview of submitted systems and final results of the competition. It is worth noting, there is a variation, or even inconsistency, in the structure and the level of detail in the description. Participants are also encouraged to submit papers with systems architecture explanations. However, there is a lack of overall analysis across different tasks and years in SemEval. There are existing studies on the analysis of specific SemEval tasks. Nakov et al. (2016) focuses on developing Sentiment Analysis tasks in 2013-2015. Sygkounas et al. (2016) is an example of a replication study of the top performing systems, in this case systems used in SemEval Twitter Sentiment Analysis (2013-2015), and focuses on architectures and performance. Evolution and challenges in semantics similarity were described in Jimenez et al. (2015). This is an example of a study on the performance of a given type of architecture across tasks of the same type. There also exist studies on shared tasks in given domain,

specially in clinical application of NLP (Filannino and Uzuner, 2018), (Chapman et al., 2011). However, they refer to tasks outside the SemEval and are more result oriented rather than task organization. Some studies discuss ethical issues in the organisation and participation of shared tasks. An overview focusing on task competitive nature and fairness can be found in Parra Escartín et al. (2017). In Nissim et al. (2017) authors also relate to these issues, yet giving the priority to advancing the field over fair competition.

Comparatively, this paper covers a wider range of NLP topics, and compares sentiment analysis and semantic similarity as well as other task types/groups in a systematic manner. To the best to our knowledge this is the first systematic analysis on SemEval.

3 Analysis methodology

We build a corpus based on the ACL anthology archive from the SemEval workshops between the years 2012-2019. Reference material included ACL anthology papers covering the task description, tasks' websites and papers describing the participating systems. All the reference papers included in this analysis are reported in the Appendix B. The pre-processing analysis consisted in manually extracting the target categories for the analysis which includes: task types, input and output types, as well as evaluation metrics, number of teams, languages and system architectures. Tasks were grouped based on the similarity between task types. If the same team took part in several tasks the same year, we considered each participation as distinct. There are four missing tasks in the plotted indexes, due to cancellation (2015-task16, 2019-task11), task-sharing (2013-task6) or lack of supporting task description (2013-task14). Numbers of citations are the numbers returned by Google Scholar, using *Publish and Perish* supporting API (Harzing, 2007). The list of citations were manually validated and noisy entries were filtered out. A final table with all the values extracted from the corpus is included in the Appendix B.

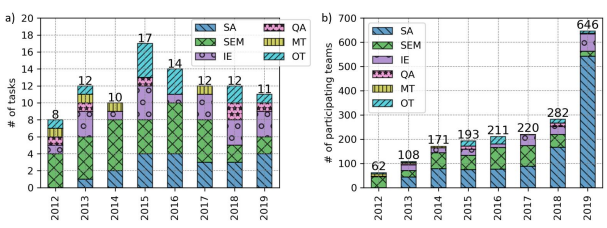

Figure 1: a) # of tasks ; b) # of teams participating in SemEval 2012-2019

4 Task types and groups

Based on task description we group each task within a macro-category. Then, due to a large number of task types, tasks were clustered within 6 groups: *Sentiment Analysis* (**SA**); *Semantic Analysis* (**SEM**): semantic analysis, semantic difference, semantic inference, semantic role labeling, semantic parsing, semantic similarity, relational similarity; *Information Extraction* (**IE**): information extraction, temporal information extraction, argument mining, fact checking; *Machine Translation* (**MT**); *Question Answering* (**QA**); *Other* (**OT**): hypernym discovery, entity linking, lexical simplification, word sense disambiguation, taxonomy extraction, taxonomy enrichment. There are also macro-categories defined by the SemEval organizers, starting from 2015, but we found them not consistent enough for the purpose of this analysis.

5 SemEval tasks in years

Within 8 editions of SemEval, a total of 96 tasks were successfully announced. The number of tasks within one group is roughly similar every year (except for MT), as well as distribution of tasks in each edition. According to Fig.1a, we observe decreasing number of SEM tasks: 5 on average in 2012-2017, and only 2 in 2018-2019. Moreover, there were no machine translation tasks in the last 2 years, and a low number of MT tasks in general (only 4 tasks in 8 years).

Although SA has a relatively limited task complexity when compared to SEM or IE, which reflects a higher variety of task types and an abundance of more specific interpretation challenges, the number of SA tasks each year is high (4, 3, 3 and 4 in years 2016-2019). It is worth mentioning, that there are other 6 SA tasks in the forthcoming SemEval 2020. The absence of some task types may be caused by the emergence of specialized workshops or conferences, e.g. low number of MT tasks in SemEval is caused by the presence a separate venue for MT: the Conference On Machine Translation (Barrault et al., 2019), which attracts more participants than SemEval in this field.

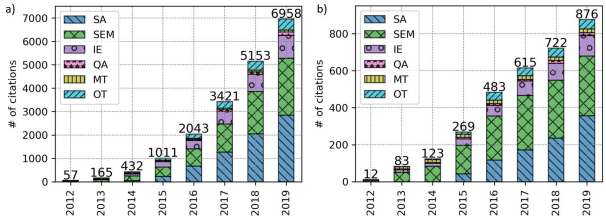

Figure 2: Cumulative number of task citations a) except for citations in SemEval proceedings; b) in SemEval proceedings

6 Task popularity

As a measure of task popularity, we analysed how many teams participated in a given task. As the number of teams signed up to the task is usually much higher than the number submitting a system result, we consider only the latter.

The number of teams increased significantly from 62 in 2012 to 646 in 2019, which shows not only a popularity boost for SemEval, but an increase in the general interest for NLP. So far, a total of 1883 teams participated in this period.

In Fig.1b, we observe a gradual increase in SemEval popularity, 30% on average each year to 2018, with a +129% jump in 2019. This is associated mainly with a dramatic increase of interest for SA: 542 teams (84% of total) in 2019. However, at the same time, number of teams in non-SA tasks decreased from 132 in 2017, to 115 in 2018 and 104 in 2019.

The most popular tasks groups along the years are SA and SEM, which gather more than 75% of teams on average each year. The third most popular is IE, in which total of 235 teams participated in SemEval from 2012 (12% of total). As a contrast, we observe a relatively low interest in QA and OT tasks. Only 41 teams participated in the last 3 years (3% of a total of 1148 in 2017-2019). Especially in OT tasks, which concentrates novel tasks, in many cases including novel formats.

In the last 2 years, SA shows a major increase in popularity (76% of all teams, compared to 40% in 2013-2017). At the same time, in tasks such as 2019-10, 2018-4 and 2018-6, which are mathematical question answering, entity linking on multiparty dialogues and parsing time normalization, respectively, only 3, 4 and 1 teams submitted results. This divergence may be associated with an emergence of easily applicable ML systems and libraries, which better fit to standard classification tasks more prevalent in SA (in contrast to OT, QA nor IE).

7 The impact of SemEval papers

As a measure of the impact of SemEval outcomes in the NLP community, we analysed the numbers of citations per task description in Google Scholar. The task description paper was used as a proxy to analyse the task impact within the NLP community. Papers submitted by participating teams describing systems and methods were not included on this analysis.

We considered the cumulative citations from 2012 to 2019 (Fig.2a), with additional distinction on citations of task description papers published in a given year (Fig.3a). Citations within SemEval proceedings were treated separately, as we focused on the impact both outside (Fig.2a) and inside (Fig.2b) the SemEval community. In other words, citations found in Google Scholar are split into numbers of papers *out* and *in* the SemEval proceedings.

SA and SEM have the highest impact, being the most cited tasks along the years both inside and outside SemEval community, what can be attributed to their high popularity.

Considering the external impact, in 2019 SA and SEM anthologies contributed with 2847 (41%) and 2426 (35%) citations respectively. IE has 985 citations (14%) and QA contributed with 148 citations (2%). The OT group, which consists of less canonical tasks, accumulated 468 citations (7%). The impact of MT papers is noticeably lower - 84 (1%).

In terms of citations within the SemEval community (in all SemEval 2012-2019 proceedings), we observe a similar pattern: 41% and 37% citations in 2019 come from SA and SEM (357 and 322), and for remaining task groups proportions are almost identical as in citations outside community (Chi.sq. *p-value*=0.06).

The number of citations outside is 8 times higher than inside the community. This proves the scientific impact and coverage, which leads to beneficial effect of SemEval on the overall NLP community.

A total of 6958 citations from 2019 are depicted in Fig.3a with distinction on the year in which the task was published (e.g. tasks from 2016 are cited 1682 times (23%)). Similarly, a total of 876 citations in the SemEval proceedings are presented in Fig.3b (e.g. anthologies published in 2015 are cited 163 times in all SemEval proceedings so far). SA tasks from 2016, SEM from 2014 and IE from 2013

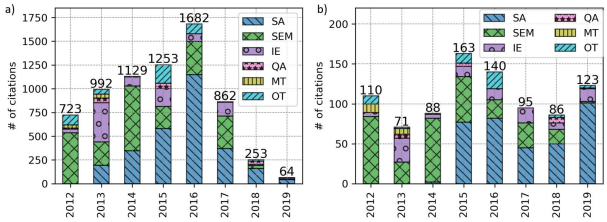

Figure 3: Number of task citations a) published in given year, except for citations in SemEval proceedings; b) from given year in SemEval proceedings

have the highest impact compared within groups (40%, 28% and 42% respectively). One could expect higher numbers of citations for older papers, however, we do not observe this pattern.

8 Languages in tasks

We analysed SemEval in terms of languages used in the tasks (Fig.4). We can distinguish 3 clusters: English-only (except for 3 tasks entirely in Chinese); multi-lingual, which define identical subtasks for several languages; cross-lingual (targeting the use of semantic relation across languages).

In total of 96 tasks, 30 investigated more than one language (multi-lingual and cross-lingual tasks) and 63 tasks were using only English.

The five most popular languages, excluding English are: Spanish (16), French (10), Italian (10), Arabic (8), German (8). Although Chinese is the 1st language in number of speakers, only 4 tasks were organised for Chinese.

Most of multi-lingual or cross-lingual tasks are related to SA (5 in 2016-2018) or SEM (15 in 2012-2019), and obviously on MT tasks (3 in 2012-2014). There were 3 OT tasks, only one QA task, and no IE tasks. Task 11 in 2017 concerning program synthesis, aiming to translate commands in English into (program) code, attracted only one team.

In 2018 and 2019 the interest in other languages is lower compared to previous years. Languages other than English were proposed only 5 and 3 times, respectively, whereas in 2016 and 2017 we observed the occurrence of respectively 10 and 14 times.

9 Input and Output Analysis

In order to better understand the evolution of the semantic complexity of the tasks, we analysed them in terms of the types used to represent input and output data in all subtasks. Based on their descriptions, we devised a list of 25 different abstract types used, then assigning each subtask the most appropriate Input and Output Types.

9.1 Types and Clusters

Taking into consideration both their complexity and purpose, we split the type list into 5 clusters: *cluster 1*: document, text, paragraph, sentence, phrase, word, number; *cluster 2*: score, score real value, score whole value, class label, probability distribution; *cluster 3*: entity, attribute, topic, tree, Directed Acyclic Graph (DAG); *cluster 4*: question, answer, query; *cluster 5*: Knowledge Base (KB), program, time interval, timeline, semantic graph, syntactic labeled sentence.

9.2 Input Types

As expected, types from *cluster 1* (sequential tokens) make up for 76% of overall input types used in all tasks (depicted in the Appendix A, Fig.A.1). Most popular input type is paragraph, for which about 60% of cases represents a tweet. The remaining 24% is split across *clusters 2, 3, 4 and 5*. A subtle divergence towards the right-hand side can be noticed, starting with 2015, driven mostly by tasks from SA and IE task groups. The most dominant Input Types from each cluster are paragraph, class label, entity, question and KB.

9.3 Output Types

As shown in Fig.5, data types from *clusters 2 and 3* are the majority in this case, accounting for 68% of used representations. Class labels are repeatedly employed, especially by SA tasks. *Cluster 1* types are constantly used across the years, fully dependent on the task types given in a certain year, 78% of them coming from SEM, IE and OT. Rarely used are typed from *clusters 4 and 5*, accounting for just 10% of the total, half of which occur in SEM tasks during 2016 and 2017 complex tasks such as Community Question Answering and Chinese Semantic Dependency Parsing. We also found a possible relation between output type and popularity. In 2012-2017 tasks where outputs were in *cluster 4 or 5*, attracted 8.3 teams per task on average, while in *clusters 1-3* 13.9 teams/task. However, despite major increase in SemEval popularity, in 2018-2019 the former attracted only 7 teams/task, and the latter 43.5 teams/task. The group with most type variety is SEM, covering types across all clusters. On the other side of the spectrum, SA has the least variety, despite it being the most popular task group. The most dominant Output Types from each cluster are

paragraph, class label, entity, answer and semantic graph.

10 Evaluation Metrics

We counted a total of 29 different evaluation metrics used in SemEval (Fig.6).

At a subtask level, the most frequent metric is F1-score, with 105 usages, followed by recall and precision, with 51 and 49 usages respectively, and accuracy, with 26 usages. F1, recall and precision are frequently jointly used, the last two playing the role of supporting break-down metrics for F1 in 95% of cases. This combination is very popular, especially for IE tasks, almost half of the use coming from this task group.

The top 5 evaluation metrics make up 84% of the total number of metrics used in all years, last 12 (almost half) being only used once. In 89% of cases when rare evaluation metrics (from Kendall's T to the right) are used, they occur in SA and SEM tasks e.g. Jaccard index in Affect in Tweets (2018) or Smatch in Meaning Representation Parsing (2016). Furthermore, 67% of the least used evaluation metrics (only used 3 times or less) appear in 2015-2017, the same period when we could see tasks experimenting the most with input and output types.

10.1 Evaluation Metrics against Output Types

F1, recall and precision (depicted in Appendix A, Fig.A.2) are mostly used for output types such as class label, paragraph and entity (each of which is the top output type from their clusters). Meanwhile, for output types represented by score, most used evaluation metrics are Pearson Correlation, Kendall's T, cosine similarity and Spearman Correlation. MAP, the 6th most used evaluation metric, is mostly used for ranked questions/answers either in recurring tasks such as Community Question Answering. Human judgment was only used twice, in Taxonomy Extraction Evaluation (2016) and Abstract Meaning Representation Parsing and Generation (2017). For further reference, see Appendix A.

11 Zooming in into Sentiment Analysis

11.1 System architectures

The systematic analysis of the prevalent methods and architectures imposed particular challenges with regard to the data extraction process due to the intrinsic complexity of tasks (many systems

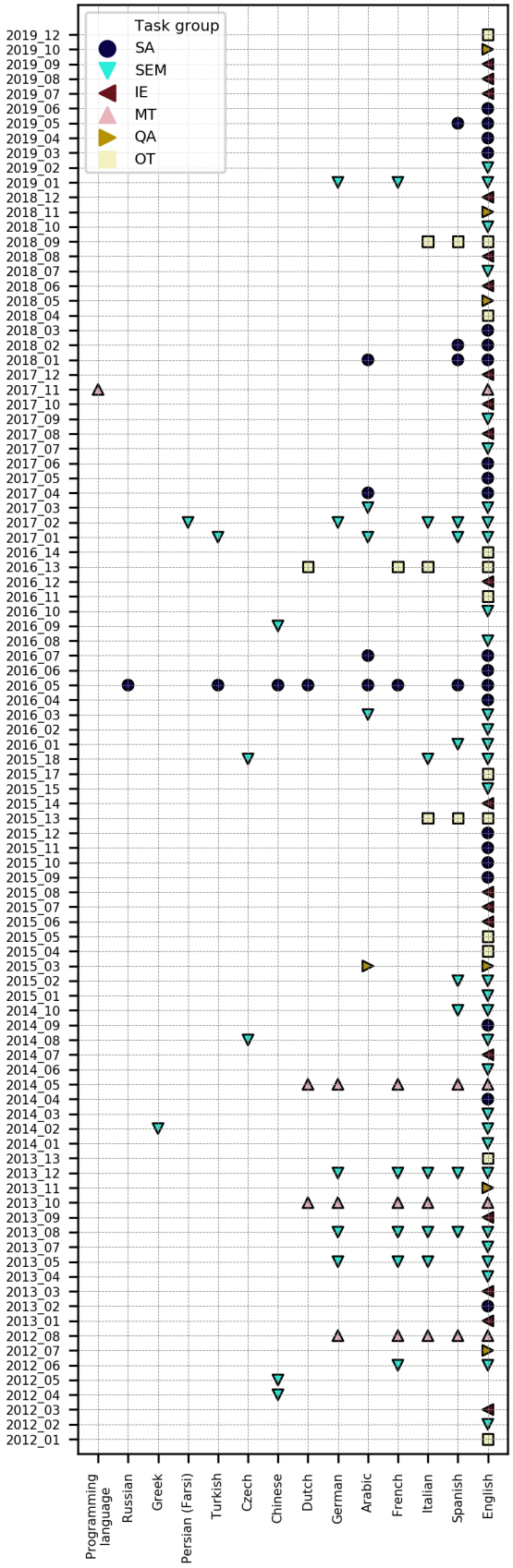

Figure 4: Languages used in SemEval tasks from 2012 to 2019

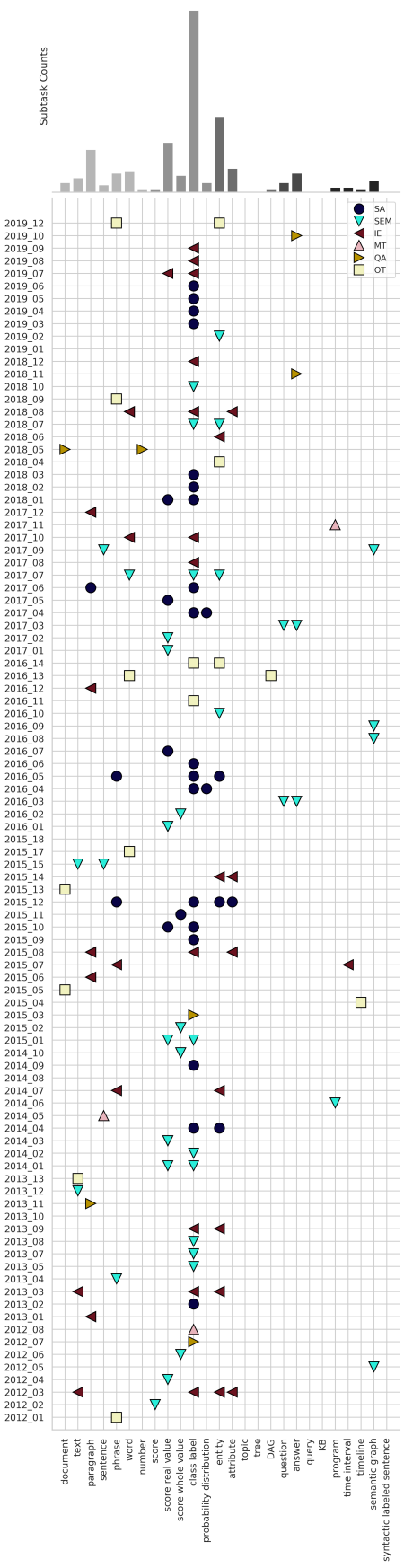

Figure 5: Output Types used in SemEval tasks from 2012 to 2019

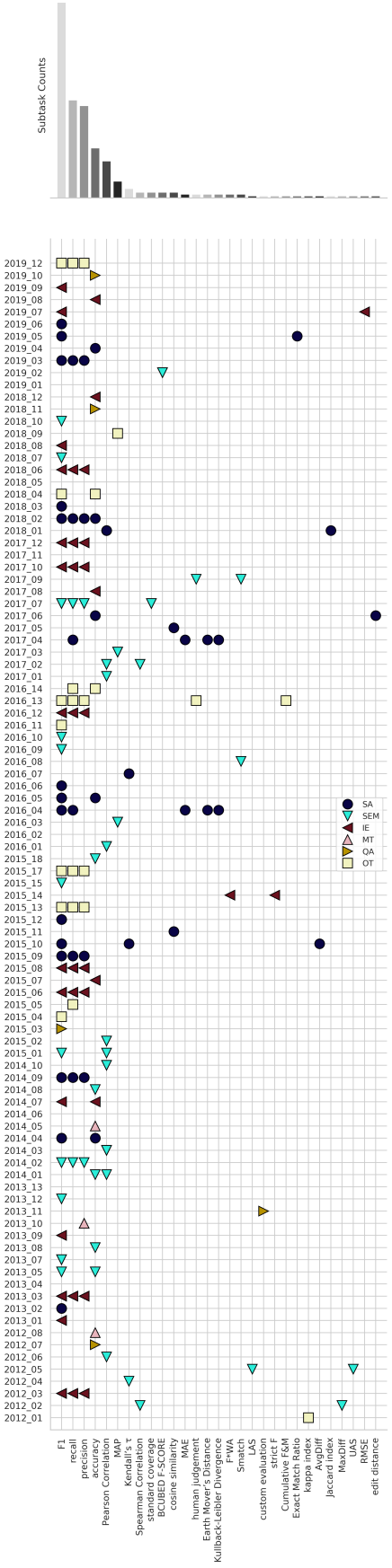

Figure 6: Evaluation Metrics used in SemEval tasks from 2012 to 2019

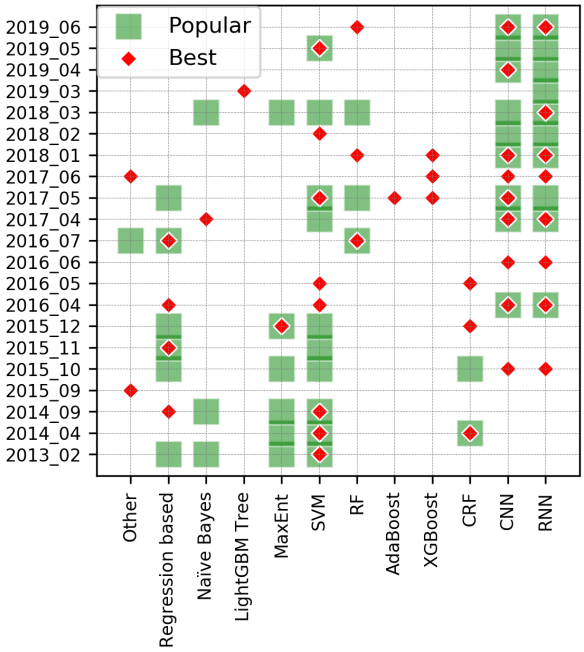

Figure 7: Models used in SA tasks from 2012 to 2019 at SemEval

include the composition of pre-processing techniques, rules, hand-crafted features and combinations of algorithms). Additionally, for the majority of task description papers, there is no systematic comparison between systems within a task, and consequently within group or years.

Due to the consistent presence of SA along all years, we present an overview of the evolution of system architectures used in SA from 2013 to 2019 (Fig.7). In this analysis we focus on the best performing architectures. More than one best model in a task signifies best models in subtasks or that the final system was an ensemble of several algorithms. *Regression based* model encompasses linear, logistic, or Gaussian regression, and *Other* includes all rule-based or heavily hand-crafted models.

We observe a drift in popularity of architectures from ML algorithms (2013-2016) to deep learning (DL) models (2017-2019).

Despite the major adoption of DL models, traditional ML algorithms are consistently in use, both as separate models and as ensembles with DL. This is also true for other types of tasks. In many task description papers from 2018-2019, one can find ML-based systems as top performing participants. SVM-based models are still popular and in some tasks outperforms DL (2018-2, 2019-5).

In the analysis of system architectures one needs to take into account that best system depends not only on the core algorithm but also on the team expertise and supporting feature sets and language resources.

11.2 Representations

The output of the SA related tasks provide an account of the evolution of sentiment and emotion representation in this community from 2013 until 2019 (see Appendix A Fig.A.3).

At a discrete level, the number of maximum class labels representing sentiment intensity grew from 3 in 2013 to 7 in 2019. At a continuous score level, real-valued scores associated with sentiment was first used in 2015; in 2016 it switched to sentiment intensity; in 2017 it was being used as a way to determine the intensity of an emotion component out of 11 emotion types (rather than a single one, or the generic emotional intensity of a sentence). In terms of targeted subject, the tasks grew more granular over time: paragraph/word (2013), aspect terms (2014), sentence topic (2015), person (2016). Additionally, discourse evolved from simpler opinionated text in the direction of figurative language, for example: handling irony and metaphor in SA (2015), phrases comparison/ranking in terms of sense of humor (2017), irony detection (2018) and contextual emphasis (2019).

12 Discussion: What is SemEval evaluating?

The results of the analysis substantiate the following core claims, which summarise some of the trends identified in this paper:

- There is evidence of significant impact of SemEval in the overall NLP community.
- SemEval contributed to the construction of a large and diverse set of challenges with regard to semantic representation, supporting resources and evaluation methodologies and metrics.
- SemEval is becoming heavily biased towards solving classification/regression problems. We observe a major interest in tasks where the expected output is a binary or multi-class label or within a continuous real valued score.
- Sentiment Analysis tasks accounts for a disproportional attention from the community.

- There are two parallel narratives running on SemEval: low entry barrier and state-of-the-art defining. SemEval contains a rich corpus of unaddressed and complex NLP tasks, which are eclipsed by the easier low entry barrier tasks. This points to the double function of SemEval which performs a pedagogical task, serving as an entry point for early career researchers to engage within the NLP community and a state-of-the-art forum for pushing the boundaries of natural language interpretation. With the popularity of NLP applications and Deep Learning, the former function is eclipsing the latter.
- There is a significant trend to decrease the variety in the output and evaluation metrics in the recent years. While in the previous years, tasks focused more on novel and exploratory tasks, recent tasks have explored, probably due to emergence of out-of-the-box DL models, this variety significantly decreased. Consequently, participants focus on easier tasks, which in part dissipates the community potential to address long-term challenges.
- Despite the recent interest in neural-based architectures, there is clear evidence of the longevity and lasting impact of older NLP methods.

13 Recommendations

We believe that this paper can serve as a guideline for the selection and organisation of future SemEval tasks. Based on the analyses performed on this paper, these are the main recommendations:

- Prioritise tasks which have a clear argument on semantic and methodological challenges and novelty.
- Differentiate challenges which have a competition/pedagogical purpose from research tasks.
- Support the systematic capture of task metadata and submission data in a structured manner. This will allow for an efficient comparison between SemEval tasks and deriving insights for future SemEval editions.

14 Conclusions

This paper reported a systematic quantitative analysis of SemEval, which is an important venue for the empirical evaluation of NLP systems. The analysis, which provides a detailed breakdown of 96 tasks in the period between 2012-2019, provided quantitative evidence that:

- SemEval has a significant impact in the overall NLP community
- there is a recent drift towards the direction of Deep Learning classification methods which is eclipsing the research function of SemEval
- there is longevity and impact of older NLP methods in comparison to Deep Learning methods

References

Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors. 2012. **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, Montréal, Canada.

Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors. 2018. *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors. 2017. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada.

Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors. 2016. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Michele Filannino and Özlem Uzuner. 2018. Advancing the state of the art in clinical natural language processing through shared tasks. *Yearbook of medical informatics*, 27(1):184—192.

A.W. Harzing. 2007. Publish or perish. Available from https://harzing.com/resources/publish-or-perish.

Sergio Jimenez, Fabio A. Gonzalez, and Alexander Gelbukh. 2015. Soft Cardinality in Semantic Text Processing: Experience of the SemEval International Competitions. *Polibits*, pages 63 – 72.

Suresh Manandhar and Deniz Yuret, editors. 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, Atlanta, Georgia, USA.

Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors. 2019. *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA.

Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M. Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2016. Developing a successful semeval task in sentiment analysis of twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

Preslav Nakov and Torsten Zesch, editors. 2014. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, Dublin, Ireland.

Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors. 2015. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado.

Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. Last words: Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Efstratios Sygkounas, Giuseppe Rizzo, and Raphaël Troncy. 2016. A replication study of the top performing systems in semeval twitter sentiment analysis. In *The Semantic Web – ISWC 2016*, pages 204–219, Cham. Springer International Publishing.

A Appendix A

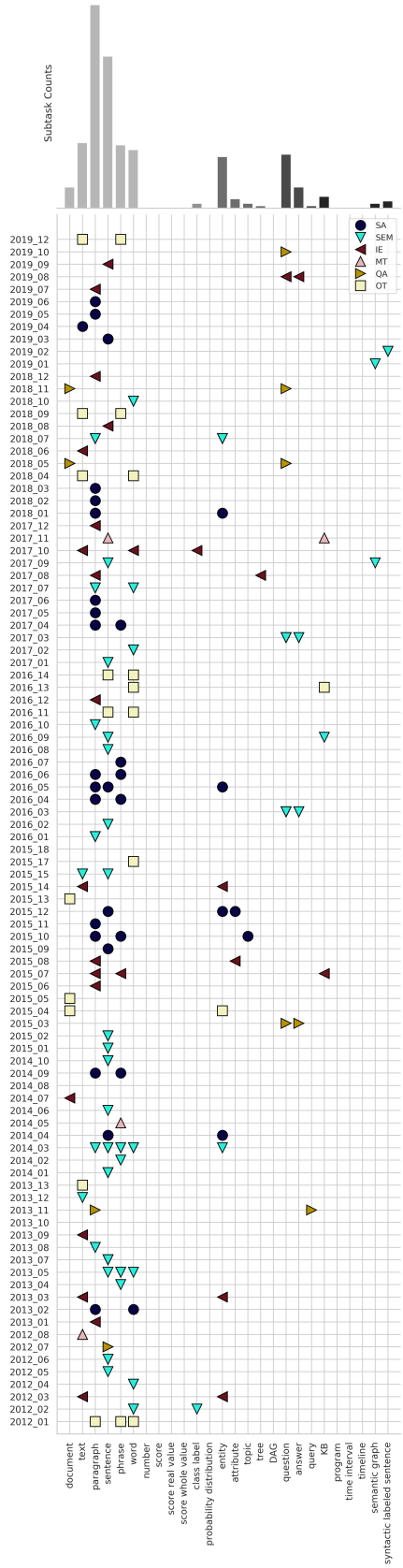

Figure A.1: Input Types used in SemEval tasks from 2012 to 2019

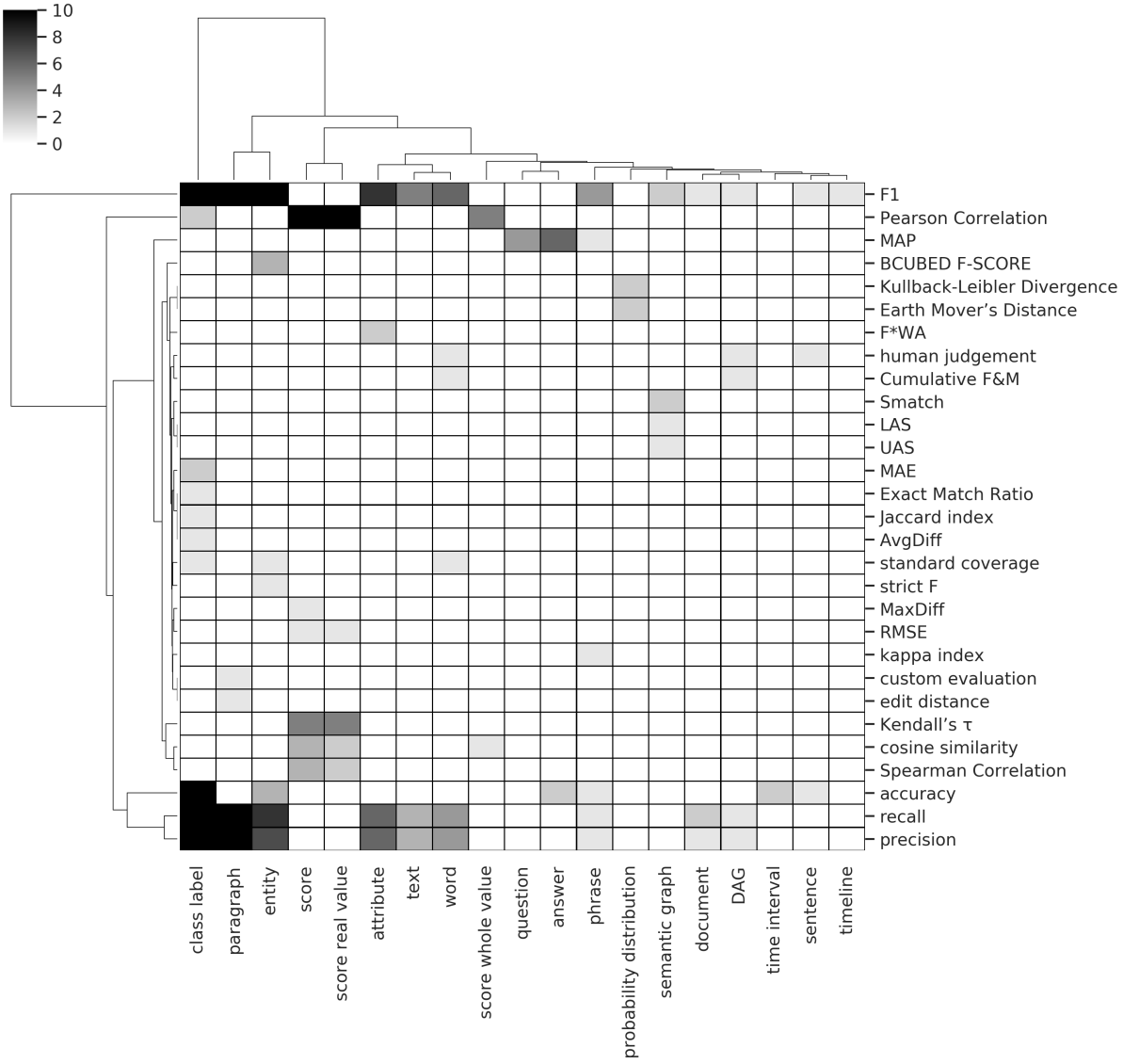

Figure A.2: Heatmap on the Evaluation Metrics and Output Types

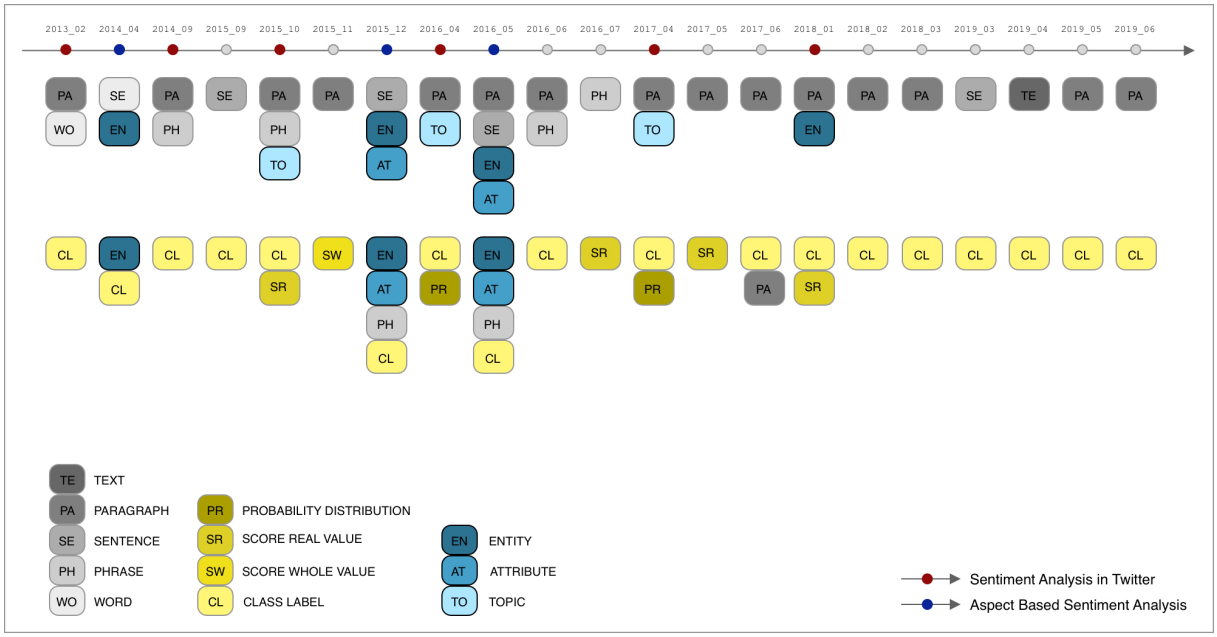

Figure A.3: Timeline of Input Types (upper row) and Output Types (lower row) in Sentiment Analysis tasks at SemEval 2013-2019

Appendix B

Task ID	Full Name	Citations	Task Type	Task Group	Input	Output	Evaluation Metric	Languages	No of Teams
2012_01	1. English Lexical Simplification[1]	103	lexical simplification	OT	word, paragraph, phrase	phrase	kappa index	English	5
2012_02	2. Measuring Degrees of Relational Similarity[2]	122	relational similarity	SEM	word, class label	score	MaxDiff, Spearman Correlation	English	3
2012_03	3. Spatial Role Labeling[3]	39	information extraction	IE	entity, text	entity, attribute, class label, text	recall, F1, precision	English	1
2012_04	4. Evaluating Chinese Word Similarity[4]	35	semantic similarity	SEM	word	score real value	Kendall's τ	Chinese	2
2012_05	5. Chinese Semantic Dependency Parsing[5]	16	semantic parsing	SEM	sentence	semantic graph	LAS, UAS	Chinese	5
2012_06	6. Semantic Textual Similarity[6]	363	semantic similarity	SEM	sentence	score whole value	Pearson Correlation	English, French	35
2012_07	7. COPA: Choice Of Plausible Alternatives An evaluation of commonsense causal reasoning[7]	11	question answering	QA	sentence	class label	accuracy	English	1
2012_08	8. Cross-lingual Textual Entailment for Content Synchronization[8]	34	machine translation	MT	text	class label	accuracy	English, Spanish, Italian, French, German	10

[1] "SemEval-2012 Task 1: English Lexical Simplification - ACL ...." https://www.aclweb.org/anthology/S12-1046. Accessed 28 May. 2020.
[2] "SemEval-2012 Task 2: Measuring Degrees of Relational ...." 7 Jun. 2012, https://www.aclweb.org/anthology/S12-1047.pdf. Accessed 28 May. 2020.
[3] "SemEval-2013 Task 3: Spatial Role Labeling - ACL Anthology." https://www.aclweb.org/anthology/S13-2044. Accessed 28 May. 2020.
[4] "SemEval-2012 Task 4: Evaluating Chinese Word Similarity ...." https://www.aclweb.org/anthology/S12-1049. Accessed 28 May. 2020.
[5] "SemEval-2012 Task 5: Chinese Semantic Dependency Parsing." https://www.aclweb.org/anthology/S12-1050. Accessed 28 May. 2020.
[6] "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity ...." https://www.aclweb.org/anthology/S12-1051. Accessed 28 May. 2020.
[7] "SemEval-2012 Task 7: Choice of Plausible Alternatives: An ...." https://www.aclweb.org/anthology/S12-1052. Accessed 28 May. 2020.
[8] "Semeval-2012 Task 8: Cross-lingual Textual Entailment for ...." https://www.aclweb.org/anthology/S12-1053. Accessed 28 May. 2020.

2013_01	1. TempEval-3 Temporal Annotation[9]	213	information extraction	IE	paragraph	paragraph	F1	English	9
2013_02	2. Sentiment Analysis in Twitter[10]	195	sentiment analysis	SA	word, paragraph	class label	F1	English	44
2013_03	3. Spatial Role Labeling[11]	29	information extraction	IE	entity, text	entity, class label, text	recall, F1, precision	English	1
2013_04	4. Free Paraphrases of Noun Compounds[12]	30	semantic similarity	SEM	phrase	phrase	?	English	3
2013_05	5. Evaluating Phrasal Semantics[13]	23	semantic inference	SEM	word, sentence, phrase	?, class label	accuracy, F1	English, Italian, French, German	5
2013_07	7. The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge[14]	76	semantic inference	SEM	sentence	class label	F1	English	9
2013_08	8. Cross-lingual Textual Entailment for Content Synchronization[15]	10	semantic inference	SEM	paragraph	class label	accuracy	English, Spanish, Italian, French, German	6
2013_09	9. Extraction of Drug-Drug Interactions from BioMedical Texts[16]	170	information extraction	IE	text	class label, entity	F1	English	14
2013_10	10. Cross-lingual Word Sense Disambiguation[17]	42	machine translation	MT	?	?	precision	English, Italian, French, German, Dutch	5
2013_11	11. Evaluating Word Sense Induction & Disambiguation within An End-User Application[18]	48	question answering	QA	query, paragraph	paragraph	custom evaluation	English	5

[9] "SemEval-2013 Task 1: TempEval-3: Evaluating Time ...." 14 Jun. 2013, https://www.aclweb.org/anthology/S13-2001.pdf. Accessed 28 May. 2020.
[10] "SemEval-2013 Task 2: Sentiment Analysis in Twitter." 14 Jun. 2013, https://www.aclweb.org/anthology/S13-2052.pdf. Accessed 28 May. 2020.
[11] "SemEval-2013 Task 3: Spatial Role Labeling - ACL Anthology." https://www.aclweb.org/anthology/S13-2044. Accessed 28 May. 2020.
[12] "SemEval-2013 Task 4: Free Paraphrases of Noun Compounds." https://www.aclweb.org/anthology/S13-2025. Accessed 28 May. 2020.
[13] "SemEval-2013 Task 5: Evaluating Phrasal Semantics - ACL ...." https://www.aclweb.org/anthology/S13-2007. Accessed 28 May. 2020.
[14] "SemEval-2013 Task 7: The Joint Student Response Analysis ...." https://www.aclweb.org/anthology/S13-2045. Accessed 28 May. 2020.
[15] "Semeval-2013 Task 8: Cross-lingual Textual Entailment ... - ACL." https://www.aclweb.org/anthology/S13-2005. Accessed 28 May. 2020.
[16] "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions ...." https://www.aclweb.org/anthology/S13-2056. Accessed 28 May. 2020.
[17] "SemEval-2013 Task 10: Cross-lingual Word Sense ... - cs.York." https://www.cs.york.ac.uk/semeval-2013/accepted/25_Paper.pdf. Accessed 28 May. 2020.
[18] "SemEval-2013 Task 11: Word Sense Induction and ...." https://www.aclweb.org/anthology/S13-2035. Accessed 28 May. 2020.

2013_12	12. Multilingual Word Sense Disambiguation[19]	107	semantic labeling	SEM	text	text	F1	English, Spanish, Italian, French, German	3
2013_13	13. Word Sense Induction for Graded and Non-Graded Senses[20]	49	word sense disambiguation	OT	text	text	?	English	4
2014_01	1. Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Entailment[21]	374	semantic similarity	SEM	sentence	score real value, class label	accuracy, Pearson Correlation	English	21
2014_02	2. Grammar Induction for Spoken Dialogue Systems [22]	0	semantic similarity	SEM	phrase	class label	recall, F1, precision	English, Greek	3
2014_03	3. Cross-Level Semantic Similarity[23]	42	semantic similarity	SEM	word, phrase, entity, paragraph, sentence	score real value	Pearson Correlation	English	19
2014_04	4. Aspect Based Sentiment Analysis[24]	344	sentiment analysis	SA	entity, sentence	class label, entity	accuracy, F1	English	32
2014_05	5. L2 Writing Assistant[25]	4	machine translation	MT	partial sentence, phrase	sentence	accuracy	English, Spanish, French, German, Dutch	6
2014_06	6. Supervised Semantic Parsing of Spatial Robot Commands[26]	15	semantic parsing	SEM	sentence	program	?	English	6
2014_07	7. Analysis of Clinical Text[27]	96	information extraction	IE	document	entity, phrase	accuracy, F1	English	21
2014_08	8. Broad-Coverage Semantic Dependency Parsing[28]	98	semantic parsing	SEM	?	?	accuracy	English, Czech	8

[19] "SemEval-2013 Task 12: Multilingual Word Sense ...." https://www.aclweb.org/anthology/S13-2040. Accessed 28 May. 2020.
[20] "SemEval-2013 Task 13: Word Sense Induction for Graded and ...." https://www.aclweb.org/anthology/S13-2049. Accessed 28 May. 2020.
[21] "SemEval-2014 Task 1: Evaluation of Compositional ...." 24 Aug. 2014, https://www.aclweb.org/anthology/S14-2001.pdf. Accessed 28 May. 2020.
[22] "Grammar Induction for Spoken Dialogue Systems - ACL ...." https://www.aclweb.org/anthology/S14-2002. Accessed 28 May. 2020.
[23] "SemEval-2014 Task 3: Cross-Level Semantic Similarity - ACL ...." https://www.aclweb.org/anthology/S14-2003. Accessed 28 May. 2020.
[24] "SemEval-2014 Task 4: Aspect Based Sentiment Analysis ...." https://www.aclweb.org/anthology/S14-2004. Accessed 28 May. 2020.
[25] "SemEval 2014 Task 5 - L2 Writing Assistant - ACL Anthology." https://www.aclweb.org/anthology/S14-2005. Accessed 28 May. 2020.
[26] "SemEval-2014 Task 6: Supervised Semantic Parsing of ...." https://www.aclweb.org/anthology/S14-2006. Accessed 28 May. 2020.
[27] "SemEval-2014 Task 7: Analysis of Clinical Text - ACL Anthology." https://www.aclweb.org/anthology/S14-2007. Accessed 28 May. 2020.
[28] "SemEval 2014 Task 8: Broad-Coverage Semantic ... - ACL." https://www.aclweb.org/anthology/S14-2008. Accessed 28 May. 2020.

2014_09	9. Sentiment Analysis in Twitter[29]	3	sentiment analysis	SA	paragraph, phrase	class label	recall, F1, precision	English	46
2014_10	10. Multilingual Semantic Textual Similarity[30]	153	semantic similarity	SEM	sentence	score whole value	Pearson Correlation	English, Spanish	9
2015_01	Task 1: Paraphrase and Semantic Similarity in Tw[31] itter	70	semantic similarity	SEM	sentence	score real value, class label	Pearson Correlation, F1	English	19
2015_02	Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability[32]	161	semantic similarity	SEM	sentence	score whole value	Pearson Correlation	English, Spanish	29
2015_03	Task 3: Answer Selection in Community Question Answering[33]	56	question answering	QA	question, answer	class label	F1	English, Arabic	13
2015_04	Task 4: TimeLine: Cross-Document Event Ordering[34]	43	temporal analysis	OT	document, entity	timeline	F1	English	4
2015_05	Task 5: QA TempEval[35]	24	temporal analysis	OT	document	document	recall	English	2
2015_06	Task 6: Clinical TempEval[36]	83	information extraction	IE	paragraph	paragraph	recall, F1, precision	English	3
2015_07	Task 7: Diachronic Text Evaluation[37]	31	information extraction	IE	paragraph, KB, phrase	time interval, phrase	accuracy	English	4
2015_08	Task 8: SpaceEval[38]	21	information extraction	IE	attribute, paragraph	attribute, class label, paragraph	recall, F1, precision	English	3
2015_09	Task 9: CLIPEval Implicit Polarity of Events[39]	13	sentiment analysis	SA	sentence	class label	recall, F1, precision	English	2
2015_10	Task 10: Sentiment Analysis in Twitter[40]	252	sentiment analysis	SA	topic, paragraph, phrase	score real value, class label	Kendall's τ, F1, AvgDiff	English	41
2015_11	Task 11: Sentiment Analysis of Figurative Language in Twitter[41]	102	sentiment analysis	SA	paragraph	score whole value	cosine similarity	English	15

[29] "SemEval-2014 Task 9: Sentiment Analysis in Twitter - ACL." 24 Aug. 2014, https://www.aclweb.org/anthology/S14-2009.pdf. Accessed 28 May. 2020.
[30] "SemEval-2014 Task 10: Multilingual Semantic Textual Similarity." https://www.aclweb.org/anthology/S14-2010. Accessed 28 May. 2020.
[31] "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in ...." https://www.aclweb.org/anthology/S15-2001. Accessed 28 May. 2020.
[32] "SemEval-2015 Task 2: Semantic Textual Similarity, English ...." https://www.aclweb.org/anthology/S15-2045. Accessed 28 May. 2020.
[33] "SemEval-2015 Task 3: Answer Selection in Community ...." https://www.aclweb.org/anthology/S15-2047. Accessed 28 May. 2020.
[34] "SemEval-2015 Task 4: TimeLine: Cross-Document Event ...." https://www.aclweb.org/anthology/S15-2132. Accessed 28 May. 2020.
[35] "SemEval-2015 Task 5: QA TempEval - Evaluating Temporal ...." https://www.aclweb.org/anthology/S15-2134. Accessed 28 May. 2020.
[36] "SemEval-2015 Task 6: Clinical TempEval - ACL Anthology." https://www.aclweb.org/anthology/S15-2136. Accessed 28 May. 2020.
[37] "SemEval 2015, Task 7: Diachronic Text Evaluation." https://www.aclweb.org/anthology/S15-2147.pdf. Accessed 28 May. 2020.
[38] "SemEval-2015 Task 8: SpaceEval - ACL Anthology." https://www.aclweb.org/anthology/S15-2149. Accessed 28 May. 2020.
[39] "SemEval-2015 Task 9: CLIPEval Implicit Polarity of Events ...." https://www.aclweb.org/anthology/S15-2077. Accessed 28 May. 2020.
[40] "SemEval-2015 Task 10: Sentiment Analysis in Twitter - ACL ...." https://www.aclweb.org/anthology/S15-2078. Accessed 28 May. 2020.
[41] "SemEval-2015 Task 11: Sentiment Analysis of Figurative ...." http://alt.qcri.org/semeval2015/task11/. Accessed 28 May. 2020.

2015_12	Task 12: Aspect Based Sentiment Analysis[42]	214	sentiment analysis	SA	?, attribute, entity, sentence	?, attribute, class label, entity, phrase	?, F1	English	16
2015_13	Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking[43]	85	entity linking	OT	document	document	recall, F1, precision	English, Spanish, Italian	9
2015_14	Task 14: Analysis of Clinical Text[44]	53	information extraction	IE	entity, text	attribute, entity	F*WA, strict F	English	16
2015_15	Task 15: A CPA Dictionary-Entry-Building Task[45]	2	semantic parsing	SEM	text, sentence	?, text, sentence	F1	English	5
2015_17	Task 17: Taxonomy Extraction Evaluation[46]	43	taxonomy extraction	OT	word	word	recall, F1, precision	English	6
2015_18	Task 18: Semantic Dependency Parsing[47]	0	semantic parsing	SEM	?	?	accuracy	English, Italian, Czech	6
2016_01	Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation[48]	150	semantic similarity	SEM	paragraph	score real value	Pearson Correlation	English, Spanish	43
2016_02	Task 2: Interpretable Semantic Textual Similarity[49]	26	semantic similarity	SEM	sentence	score whole value	?	English	9
2016_03	Task 3: Community Question Answering[50]	123	semantic similarity	SEM	question, answer	question, answer	MAP	English, Arabic	18
2016_04	Task 4: Sentiment Analysis in Twitter[51]	436	sentiment analysis	SA	paragraph, phrase	probability distribution, class label	MAE, Kullback-Leibler Divergence, recall, F1, Earth Mover's Distance	English	14
2016_05	Task 5: Aspect-Based Sentiment Analysis[52]	465	sentiment analysis	SA	?, paragraph, entity, sentence	?, class label, entity, phrase	accuracy, F1	English, Spanish, French, Arabic, Dutch, Chinese,	29

[42] "SemEval-2015 Task 12: Aspect Based Sentiment Analysis ...." https://www.aclweb.org/anthology/S15-2082/. Accessed 28 May. 2020.
[43] "SemEval-2015 Task 13: Multilingual All-Words Sense ... - ACL." https://www.aclweb.org/anthology/S15-2049. Accessed 28 May. 2020.
[44] "SemEval-2015 Task 14: Analysis of Clinical Text - ACL ...." https://www.aclweb.org/anthology/S15-2051. Accessed 28 May. 2020.
[45] "SemEval-2015 Task 15: A CPA dictionary-entry-building task." https://www.aclweb.org/anthology/S15-2053.pdf. Accessed 28 May. 2020.
[46] "SemEval-2015 Task 17: Taxonomy Extraction Evaluation ...." https://www.aclweb.org/anthology/S15-2151. Accessed 28 May. 2020.
[47] "SemEval 2015 Task 18: Broad-Coverage Semantic ...." https://www.aclweb.org/anthology/S15-2153. Accessed 28 May. 2020.
[48] "SemEval-2016 Task 1: Semantic Textual Similarity ...." https://www.aclweb.org/anthology/S16-1081. Accessed 28 May. 2020.
[49] "SemEval-2016 Task 2: Interpretable Semantic Textual Similarity." https://www.aclweb.org/anthology/S16-1082. Accessed 28 May. 2020.
[50] "SemEval-2016 Task 3: Community Question Answering - ACL ...." https://www.aclweb.org/anthology/S16-1083. Accessed 28 May. 2020.
[51] "SemEval-2016 Task 4: Sentiment Analysis in Twitter - ACL ...." https://www.aclweb.org/anthology/S16-1001. Accessed 28 May. 2020.
[52] "SemEval-2016 Task 5: Aspect Based Sentiment Analysis ...." https://www.aclweb.org/anthology/S16-1002. Accessed 28 May. 2020.

								Turkish, Russian	
2016_06	Task 6: Detecting Stance in Tweets[53]	209	sentiment analysis	SA	paragraph, phrase	class label	F1	English	28
2016_07	Task 7: Determining Sentiment Intensity of English and Arabic Phrases[54]	39	sentiment analysis	SA	phrase	score real value	Kendall's τ	English, Arabic	5
2016_08	Task 8: Meaning Representation Parsing[55]	28	semantic parsing	SEM	sentence	semantic graph	Smatch	English	11
2016_09	Task 9: Chinese Semantic Dependency Parsing[56]	0	semantic parsing	SEM	KB, sentence	semantic graph	F1	Chinese	3
2016_10	Task 10: Detecting Minimal Semantic Units and their Meanings[57]	23	semantic analysis	SEM	paragraph	entity	F1	English	6
2016_11	Task 11: Complex Word Identification[58]	40	lexical simplification	OT	word, sentence	class label	F1	English	21
2016_12	Task 12: Clinical TempEval[59]	79	information extraction	IE	paragraph	paragraph	recall, F1, precision	English	14
2016_13	Task 13: Taxonomy Extraction Evaluation (TExEval-2)[60]	54	taxonomy extraction	OT	word, KB	word, DAG	human judgement, recall, F1, Cumulative F&M, precision	English, Italian, French, Dutch	5
2016_14	Task 14: Semantic Taxonomy Enrichment[61]	10	taxonomy enrichment	OT	word, sentence	class label, entity	accuracy, recall	English	5
2017_01	Task 1: Semantic Textual Similarity[62]	144	semantic similarity	SEM	sentence	score real value	Pearson Correlation	English, Spanish, Arabic, Turkish	31
2017_02	Task 2: Multi-lingual and Cross-lingual Semantic Word Similarity[63]	56	semantic similarity	SEM	word	score real value	Pearson Correlation, Spearman Correlation	English, Spanish, Italian, German,	17

[53] "SemEval-2016 Task 6: Detecting Stance in Tweets - ACL ...." https://www.aclweb.org/anthology/S16-1003. Accessed 28 May. 2020.
[54] "SemEval-2016 Task 7: Determining Sentiment Intensity ... - ACL." https://www.aclweb.org/anthology/S16-1004. Accessed 28 May. 2020.
[55] "SemEval-2016 Task 8: Meaning Representation Parsing ...." https://www.aclweb.org/anthology/S16-1166. Accessed 28 May. 2020.
[56] "SemEval-2016 Task 9: Chinese Semantic Dependency Parsing." https://www.aclweb.org/anthology/S16-1167. Accessed 28 May. 2020.
[57] "SemEval-2016 Task 10: Detecting Minimal Semantic Units ...." https://www.aclweb.org/anthology/S16-1084. Accessed 28 May. 2020.
[58] "SemEval 2016 Task 11: Complex Word Identification - ACL ...." https://www.aclweb.org/anthology/S16-1085. Accessed 28 May. 2020.
[59] "SemEval-2017 Task 12: Clinical TempEval - ACL Anthology." https://www.aclweb.org/anthology/S17-2093. Accessed 28 May. 2020.
[60] "SemEval-2016 Task 13: Taxonomy Extraction Evaluation ...." https://www.aclweb.org/anthology/S16-1168. Accessed 28 May. 2020.
[61] "SemEval-2016 Task 14: Semantic Taxonomy Enrichment ...." https://www.aclweb.org/anthology/S16-1169. Accessed 28 May. 2020.
[62] "SemEval-2017 Task 1: Semantic Textual Similarity ...." https://www.aclweb.org/anthology/S17-2001. Accessed 28 May. 2020.
[63] "SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic." https://www.aclweb.org/anthology/S17-2002. Accessed 28 May. 2020.

								Persian (Farsi)	
2017_03	Task 3: Community Question Answering[64]	106	semantic similarity	SEM	question, answer	question, answer	MAP	English, Arabic	23
2017_04	Task 4: Sentiment Analysis in Twitter[65]	307	sentiment analysis	SA	paragraph, phrase	probability distribution, class label	MAE, Kullback-Leibler Divergence, recall, Earth Mover's Distance	English, Arabic	48
2017_05	Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News[66]	48	sentiment analysis	SA	paragraph	score real value	cosine similarity	English	32
2017_06	Task 6: #HashtagWars: Learning a Sense of Humor[67]	15	sentiment analysis	SA	paragraph	class label, paragraph	accuracy, edit distance	English	8
2017_07	Task 7: Detection and Interpretation of English Puns [68]	23	semantic parsing	SEM	word, paragraph	word, class label, entity	standard coverage, recall, F1, precision	English	10
2017_08	Task 8: RumourEval: Determining rumour veracity and support for rumours[69]	71	fact checking	IE	tree, paragraph	class label	accuracy	English	8
2017_09	Task 9: Abstract Meaning Representation Parsing and Generation[70]	15	semantic parsing	SEM	semantic graph, sentence	semantic graph, sentence	Smatch, human judgement	English	5
2017_10	Task 10: Extracting Keyphrases and Relations from Scientific Publications[71]	73	information extraction	IE	word, class label, text	word, class label	recall, F1, precision	English	26
2017_11	Task 11: End-User Development using Natural Language[72]	4	machine translation	MT	KB, sentence	program	?	English, Programing language	1
2017_12	Task 12: Clinical TempEval[73]	0	information extraction	IE	paragraph	paragraph	recall, F1, precision	English	11
2018_01	Task 1: Affect in Tweets[74]	98	sentiment analysis	SA	entity, paragraph	score real value, class label	Pearson Correlation, Jaccard index	English, Spanish, Arabic	75

[64] "SemEval-2016 Task 3: Community Question Answering - ACL ...." https://www.aclweb.org/anthology/S16-1083. Accessed 28 May. 2020.
[65] "SemEval-2016 Task 4: Sentiment Analysis in Twitter - ACL ...." https://www.aclweb.org/anthology/S16-1001. Accessed 28 May. 2020.
[66] "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on ...." https://www.aclweb.org/anthology/S17-2089/. Accessed 28 May. 2020.
[67] "SemEval-2017 Task 6: #HashtagWars: Learning a Sense of ...." https://www.aclweb.org/anthology/S17-2004. Accessed 28 May. 2020.
[68] "SemEval-2017 Task 7: Detection and Interpretation of English ...." https://www.aclweb.org/anthology/S17-2005. Accessed 28 May. 2020.
[69] "SemEval-2017 Task 8: RumourEval: Determining rumour ...." https://www.aclweb.org/anthology/S17-2006. Accessed 28 May. 2020.
[70] "SemEval-2017 Task 9: Abstract Meaning Representation ...." https://www.aclweb.org/anthology/S17-2090. Accessed 28 May. 2020.
[71] "SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases ...." https://www.aclweb.org/anthology/S17-2091. Accessed 28 May. 2020.
[72] "End-User Development using Natural Language - ACL ...." https://www.aclweb.org/anthology/S17-2092. Accessed 28 May. 2020.
[73] "SemEval-2017 Task 12: Clinical TempEval - ACL Anthology." https://www.aclweb.org/anthology/S17-2093. Accessed 28 May. 2020.
[74] "SemEval-2018 Task 1: Affect in Tweets - ACL Anthology." https://www.aclweb.org/anthology/S18-1001. Accessed 28 May. 2020.

2018_02	Task 2: Multilingual Emoji Prediction[75]	20	sentiment analysis	SA	paragraph	class label	accuracy, recall, F1, precision	English, Spanish	49
2018_03	Task 3: Irony Detection in English Tweets[76]	43	sentiment analysis	SA	paragraph	class label	F1	English	43
2018_04	Task 4: Character Identification on Multiparty Dialogues[77]	2	entity linking	OT	word, text	entity	accuracy, F1	English	4
2018_05	Task 5: Counting Events and Participants within Highly Ambiguous Data covering a very long tail[78]	0	question answering	QA	document, question	number, document	?	English	4
2018_06	Task 6: Parsing Time Normalizations[79]	3	information extraction	IE	text	?, entity	recall, F1, precision	English	1
2018_07	Task 7: Semantic Relation Extraction and Classification in Scientific Papers[80]	26	semantic parsing	SEM	entity, paragraph	class label, entity	F1	English	32
2018_08	Task 8: Semantic Extraction from CybersecUrity REports using Natural Language Processing (SecureNLP)[81]	3	information extraction	IE	sentence	word, attribute, class label	F1	English	9
2018_09	Task 9: Hypernym Discovery[82]	13	hypernym discovery	OT	text, phrase	phrase	MAP	English, Spanish, Italian	11
2018_10	Task 10: Capturing Discriminative Attributes[83]	5	semantic difference	SEM	word	class label	F1	English	21
2018_11	Task 11: Machine Comprehension using Commonsense Knowledge[84]	33	question answering	QA	document, question	answer	accuracy	English	11
2018_12	Task 12: Argument Reasoning Comprehension Task[85]	7	argument reasoning	IE	paragraph	class label	accuracy	English	22
2019_01	Task 1: Cross-lingual Semantic Parsing with UCCA[86]	3	semantic parsing	SEM	semantic graph	?	?	English, French, German	8

[75] "SemEval 2018 Task 2: Multilingual Emoji Prediction - ACL ...." https://www.aclweb.org/anthology/S18-1003. Accessed 28 May. 2020.
[76] "SemEval-2018 Task 3: Irony Detection in English Tweets ...." https://www.aclweb.org/anthology/S18-1005. Accessed 28 May. 2020.
[77] "SemEval 2018 Task 4: Character Identification on Multiparty ...." https://www.aclweb.org/anthology/S18-1007. Accessed 28 May. 2020.
[78] "SemEval-2018 Task 5: Counting Events and Participants in ...." 5 Jun. 2018, https://www.aclweb.org/anthology/S18-1009.pdf. Accessed 28 May. 2020.
[79] "SemEval 2018 Task 6: Parsing Time Normalizations - ACL ...." https://www.aclweb.org/anthology/S18-1011. Accessed 28 May. 2020.
[80] "SemEval-2018 Task 7: Semantic Relation Extraction and ...." https://www.aclweb.org/anthology/S18-1111. Accessed 28 May. 2020.
[81] "SemEval-2018 Task 8: Semantic Extraction from ... - ACL." https://www.aclweb.org/anthology/S18-1113. Accessed 28 May. 2020.
[82] "SemEval-2018 Task 9: Hypernym Discovery - ACL Anthology." https://www.aclweb.org/anthology/S18-1115. Accessed 28 May. 2020.
[83] "SemEval-2018 Task 10: Capturing Discriminative Attributes ...." https://www.aclweb.org/anthology/S18-1117. Accessed 28 May. 2020.
[84] "SemEval-2018 Task 11: Machine Comprehension Using - ACL." https://www.aclweb.org/anthology/S18-1119. Accessed 28 May. 2020.
[85] "SemEval-2018 Task 12: The Argument Reasoning ... - ACL." https://www.aclweb.org/anthology/S18-1121. Accessed 28 May. 2020.
[86] "SemEval-2019 Task 1: Cross-lingual Semantic Parsing with ...." https://www.aclweb.org/anthology/S19-2001. Accessed 28 May. 2020.

2019_02	Task 2: Unsupervised Lexical Semantic Frame Induction[87]	1	semantic labeling	SEM	syntactic labeled sentence	entity	BCUBED F-SCORE	English	13
2019_03	Task 3: EmoContext: Contextual Emotion Detection in Text[88]	11	sentiment analysis	SA	sentence	class label	recall, F1, precision	English	311
2019_04	Task 4: Hyperpartisan News Detection[89]	5	sentiment analysis	SA	text	class label	accuracy	English	42
2019_05	Task 5: HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter[90]	15	sentiment analysis	SA	paragraph	class label	F1, Exact Match Ratio	English, Spanish	74
2019_06	Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media[91]	13	sentiment analysis	SA	paragraph	class label	F1	English	115
2019_07	Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours[92]	7	fact checking	IE	paragraph	score real value, class label	RMSE, F1	English	22
2019_08	Task 8: Fact Checking in Community Question Answering Forums[93]	4	fact checking	IE	question, answer	class label	accuracy	English	17
2019_09	Task 9: Suggestion Mining from Online Reviews and Forums[94]	3	information extraction	IE	sentence	class label	F1	English	33
2019_10	Task 10: Math Question Answering[95]	0	question answering	QA	question	answer	accuracy	English	3
2019_12	Task 12: Toponym Resolution in Scientific Papers[96]	2	entity linking	OT	text, phrase	entity, phrase	recall, F1, precision	English	8

[87] "SemEval-2019 Task 2: Unsupervised Lexical Frame Induction ...." https://www.aclweb.org/anthology/S19-2003. Accessed 28 May. 2020.
[88] "SemEval-2019 Task 3: EmoContext Contextual Emotion ...." https://www.aclweb.org/anthology/S19-2005. Accessed 28 May. 2020.
[89] "SemEval-2019 Task 4: Hyperpartisan News Detection - ACL ...." https://www.aclweb.org/anthology/S19-2145. Accessed 28 May. 2020.
[90] "SemEval-2019 Task 5: Multilingual Detection of Hate Speech ...." https://www.aclweb.org/anthology/S19-2007. Accessed 3 Jun. 2020.
[91] "SemEval-2019 Task 6: Identifying and Categorizing Offensive ...." https://www.aclweb.org/anthology/S19-2010. Accessed 28 May. 2020.
[92] "SemEval-2019 Task 7: RumourEval, Determining Rumour ...." https://www.aclweb.org/anthology/S19-2147. Accessed 28 May. 2020.
[93] "SemEval-2019 Task 8: Fact Checking in Community Question ...." https://www.aclweb.org/anthology/S19-2149. Accessed 28 May. 2020.
[94] "SemEval-2019 Task 9: Suggestion Mining from Online ...." https://www.aclweb.org/anthology/S19-2151. Accessed 28 May. 2020.
[95] "SemEval-2019 Task 10: Math Question Answering - ACL ...." https://www.aclweb.org/anthology/S19-2153. Accessed 28 May. 2020.
[96] "SemEval-2019 Task 12: Toponym Resolution in Scientific ...." https://www.aclweb.org/anthology/S19-2155. Accessed 28 May. 2020.