# NTNU at SemEval-2018 Task 7: Classifier Ensembling for Semantic Relation Identification and Classification in Scientific Papers

**Biswanath Barik**[1], **Utpal Kumar Sikdar**[2] **and Björn Gambäck**[1]
[1]Department of Computer Science, NTNU, Norway
[2]Flytxt, Thiruvananthapuram, India
{biswanath.barik,gamback}@ntnu.no
utpal.sikdar@gmail.com

## Abstract

The paper presents NTNU's contribution to SemEval-2018 Task 7 on relation identification and classification. The class weights and parameters of five alternative supervised classifiers were optimized through grid search and cross-validation. The outputs of the classifiers were combined through voting for the final prediction. A wide variety of features were explored, with the most informative identified by feature selection. The best setting achieved $F_1$ scores of 47.4% and 66.0% in the relation classification subtasks 1.1 and 1.2. For relation identification and classification in subtask 2, it achieved $F_1$ scores of 33.9% and 17.0%,

## 1 Introduction

Scientific papers are valuable knowledge sources providing authentic insights about certain aspects of the research domains. With the advancement of scientific research, a massive growth of published articles are observed. As per the American Journal Experts (AJE) scholarly publishing report[1], approximately 2.2 million articles were added to the literature in 2016 only. The sheer volume of the ever increasing literature of any scientific discipline makes it hard for human capability and expertise to quickly process and identify information of interest. Therefore, there is a need to efficiently exploit automatic means of accessing this reliable unstructured knowledge repository.

Semantic relation extraction is one of the main information extraction tasks, and aims to identify a pair of arguments connected by certain predefined relation types based on a target application. The relation arguments are of different types such as Named Entities (Freitas et al., 2009), nominals (Hendrickx et al., 2009), general keyphrases (Gábor et al., 2016; Augenstein et al., 2017), quantitative variables (Marsi et al., 2014) or events (Barik et al., 2017), and are syntactically represented by noun phrases, clauses or larger complex structures. A semantic relation may be either symmetric (undirected) or asymmetric (hierarchical).

Supervised machine learning approaches have been successfully used for identifying semantic relations encoded in texts. Broadly, three types of supervised approaches to relation extraction have been investigated: *feature-based* (Kambhatla, 2004; Jiang and Zhai, 2007), *kernel-based* (Zelenko et al., 2003), and *neural network based* (Zeng et al., 2014; Miwa and Bansal, 2016).

In this work, various relation identification and classification subtasks of SemEval 2018 Task 7 (Gábor et al., 2018) were addressed using feature-based approaches. A wide variety of features was explored, including lexical (e.g., bag-of-words, lemmata, n-grams), syntactic (e.g., part-of-speech, parsing information), semantic (e.g., dependency information, WordNet (Miller, 1995)), and other binary indicators. A $\chi^2$-based feature selection technique was used to identify informative features. The class weights and parameters of five different classifiers—Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Multinomial Naïve Bayes (MNB), and k-Nearest Neighbor (kNN)—were optimized for each subtask through grid search and $k$-fold cross-validation. These classifiers were chosen as they are effective in identifying and classifying semantic relations in feature-based classification scenario (Barik and Marsi, 2017). The trained classifiers were ensembled using majority class labels (hard voting) for the final predictions. All classifier, feature selection and classifier ensembling modules used were implemented in the `scikit-learn` (Pedregosa et al., 2011) machine learning library.

[1]https://www.aje.com/en/arc/dist/docs/International-scholarly-publishing-report-2016.pdf



*Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 858–862
New Orleans, Louisiana, June 5–6, 2018. 

| Relation Type | Data Set | Frequency/ Percentage | | | |
|---|---|---|---|---|---|
| | | Total | (%) | Fwd % | Rev % |
| USAGE | $D_1$ | 483 | 39.33 | 61.28 | 38.72 |
| | $D_2$ | 470 | 37.67 | 68.72 | 31.28 |
| RESULT | $D_1$ | 72 | 5.86 | 72.22 | 27.78 |
| | $D_2$ | 123 | 9.86 | 69.10 | 30.90 |
| MODEL-FEATURE | $D_1$ | 326 | 26.55 | 69.32 | 30.68 |
| | $D_2$ | 175 | 14.02 | 70.29 | 29.71 |
| PART_WHOLE | $D_1$ | 234 | 19.05 | 67.52 | 32.48 |
| | $D_2$ | 196 | 15.70 | 59.70 | 40.30 |
| TOPIC | $D_1$ | 18 | 1.46 | 44.44 | 55.56 |
| | $D_2$ | 243 | 19.47 | 94.65 | 5.35 |
| COMPARE | $D_1$ | 95 | 7.74 | 100 | |
| | $D_2$ | 41 | 3.28 | 100 | |
| Total | $D_1$ | 1228 | 100 | 68.00 | 32.00 |
| | $D_2$ | 1248 | 100 | 73.63 | 26.37 |

Table 1: Relation type statistics in datasets $D_1$ and $D_2$

The tasks and the datasets are described in Section 2, while Section 3 outlines the experimental setup, system architecture and parameter optimisation. Section 4 discusses the results of the final evaluation of SemEval 2018 Task 7, where the system achieved 47.4% and 66.0% $F_1$ scores on the relation classification subtasks 1.1 and 1.2. In subtask 2, the system reached 33.9% and 17.0% $F_1$ scores for relation identification and relation classification, respectively. These results are eloborated on in Section 5, before Section 6 concludes and points to future research.

## 2 Task and Dataset Description

SemEval 2018 Task 7 (Gábor et al., 2018) consisted of two main relation extraction subtasks:

(a) identifying entity mentions related with any predefined set of relation (Subtask 2), and

(b) classifying them into specific relation types (Subtasks 1.1, 1.2, and 2).

There are six relation types, among which USAGE, RESULT, MODEL-FEATURE, PART_WHOLE, and TOPIC are asymmetric, while COMPARE is the only symmetric relation. All the relations are intra-sentential and there are no referring expressions.

The training dataset consisted of two subsets:

$D_1$: 350 abstracts of scientific papers that have been manually annotated with entity mentions and relation labels (*clean data*), and

$D_2$: 350 abstracts with entity mentions automatically labelled, but with the relations labelled manually (*noisy data*).

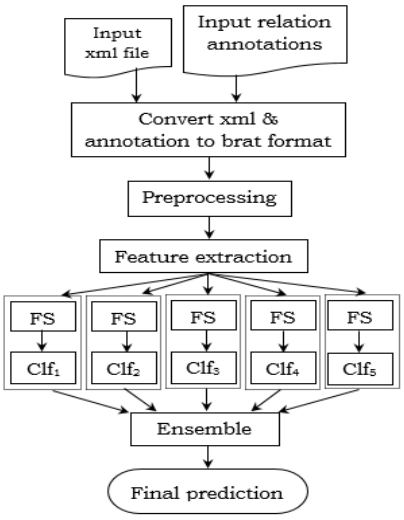


Figure 1: Relation detection and classification pipeline: 5 classifiers ($Clf_n$) work on extracted feature sets (FS)

Subtask 1.1 and subtask 2 are associated with the clean dataset $D_1$, while subtask 1.2 is associated with the noisy $D_2$ dataset. The test data consisted of 150 abstracts each for subtask 1.1, 1.2 and 2.

Table 1 shows the distribution of relation instances into different relation types, and their forward (Fwd) and reverse (Rev) directionalities in datasets $D_1$ and $D_2$. The highest number of instances are of the USAGE type in both datasets, whereas TOPIC is the least frequent relation type (1.46%) in $D_1$, but significant (19.47%) in $D_2$. The overall forward directionalities of relations are 68% in $D_1$ and 73.63% in $D_2$. The directionalities of individual relation types are similar.

The most frequent lengths of the entity mentions are two and one word(s) in $D_1$ and $D_2$, respectively, with maximum lengths of 13 and 4. The most frequent context lengths of the relation instances are two words, with a highest length of 31 words (RESULT(I05-3022.6, I05-3022.16)) in $D_1$ and 24 words (USAGE(E91-1004.30, E91-1004.37)) in $D_2$. The average number of entities in the sentences are 3 and 6 in $D_1$ and $D_2$, with highest number of entities being 17 and 29, respectively.

## 3 Experimental Setup

Figure 1 shows the processing pipeline common to both relation identification and classification. The processing steps are elaborated on below.

**Inputs to brat annotation:** The input training and test files are in `xml` format with the entity mentions marked. Each entity mention has an `ID` with two parts, abstract ID and entity number. For example, the entity ID `H91-1045.18` denotes abstract ID `H91-1045` and entity number `18`. The relation labels are in a separate file with the format `TOPIC(A92-1023.7,A92-1023.8,REVERSE)`, where the first two arguments of the relation type are entity IDs and the last is the directionality of the relation. The xml and relation label files were converted into 'brat' (Stenetorp et al., 2012) format, with the text content of each abstract ID kept in a text file, and entity and relation information kept in an annotation file. Conversion to brat format helps to visualize and study the annotations of the training and test set output. Also, the text content (without entity tags) is used for preprocessing.

**Text Processing:** The text content of each abstract is analyzed with the Stanford CoreNLP toolkit (Manning et al., 2014) for sentence boundary detection, tokenization, lemmatization, part-of-speech (POS) tagging, and constituent and dependency parsing. Character offset-based brat entity annotations are mapped into word level indices using the tokens' character offsets. Finally, the dependency heads of entity mentions, in between context and the text window representing the relation expression are identified.

**Feature Extraction:** Given a sentence with more than one entity mention, all possible entity pairs are considered in left to right order. For each entity pair, the text span containing the entities and their middle context is considered as the representation of the relation instance. As word features, unigrams and bigrams of the context and entity mentions (excluding articles, adjectives, cardinals, ordinals, pronouns, brackets and punctuations) are considered. Corresponding to word features, POS, word+POS, and lemma+POS combinations are included, as well as word and POS of entity dependency heads, context dependency heads, and their combinations.

As the shortest dependency path between the entity pair contains major information for relation identification (Bunescu and Mooney, 2005), dependency path features are added for the distance from left entity head to right entity head, words belonging to the dependency path and their relations to the parent node. WordNet synonyms and hyponyms of dependency head of entities and contexts are included. Also, other binary indicators such as adjacent or overlapping entities are included.

**Parameters Optimization through Cross-Validation (CV):** As there was no development data available for model parameter tuning, 20% of the training data was kept as development data, and the remaining training data was used for parameters optimization with 5-fold cross-validation. For relation labeling in subtask 1.1 and 1.2, the relation type is predicted against 11 classes (five directed and one undirected relation). Relation instance identification in subtask 2 is a binary classification problem, and the class weights of positive instances are optimized through CV. In the final system, the parameters are optimized on the entire training set.

**Classifiers Ensembling and Final Prediction:** The optimized parameters of the classifiers and class weights are set to the classifiers. For each classifier, the $\chi^2$-based `SelectKBest()` method selects the top $k$ features from the input feature space, where the $k$ for each classifier is determined through cross-validation. The predictions of the classifiers are then ensembled with (majority) voting where each participating classifier uses its own feature selection method.

## 4 Results

Three separate submissions were made on the test data. The first two submissions were on relation classification on clean (subtask 1.1) and noisy data (subtask 1.2). The third submission (subtask 2) consisted of relation identification followed by classification on clean data. In subtask 2, a separate system was created for the relation identification, while the relation classification system of subtask 1.1 was used for the classification.

Table 2 shows the performance (precision, recall and $F_1$ score) of individual classifiers, as well as their combinations in the relation classification subtask 1.1, where the scores are micro-averaged over all (11) classes. Among the individual classifiers, SVM gives the best result (56% $F_1$ score). Voting with the top-3 classifiers (SVM, DT & MNB) gave a slightly higher $F_1$ score of 58%.

Table 3 shows the scores of the relation classification subtask on noisy training data (subtask 1.2).

| Classifier | #Features | P | R | $F_1$ |
|---|---|---|---|---|
| SVM | 13,200 | 0.59 | 0.56 | **0.56** |
| DT | 3,400 | 0.58 | 0.53 | 0.54 |
| RF | 6,600 | 0.38 | 0.42 | 0.40 |
| MNB | 2,300 | 0.40 | 0.79 | 0.53 |
| kNN | 7,800 | 0.43 | 0.30 | 0.35 |
| Ensemble-all | – | 0.56 | 0.42 | 0.48 |
| Ensemble-best | – | 0.57 | 0.63 | **0.58** |

Table 2: Precision, recall and F-scores of individual and ensemble classifiers on subtask 1.1. The scores are micro-averaged over 11 classes. Ensemble-best is SVM+DT+MNB.

| Classifier | #Features | P | R | $F_1$ |
|---|---|---|---|---|
| SVM | 9,700 | 0.71 | 0.70 | **0.69** |
| DT | 7,200 | 0.72 | 0.66 | 0.65 |
| RF | 8,900 | 0.60 | 0.58 | 0.53 |
| MNB | 3,700 | 0.70 | 0.67 | 0.62 |
| kNN | 6,700 | 0.48 | 0.70 | 0.57 |
| Ensemble-all | – | 0.57 | 0.71 | 0.63 |
| Ensemble-best | – | 0.81 | 0.67 | **0.73** |

Table 3: Result of individual and ensemble classifiers on subtask 1.2. Scores are micro-averaged over 11 classes. Ensemble-best is SVM+RF+MNB.

As individual classifier, SVM gave the best performance with 69% $F_1$ score followed by Decision Trees (65%) and Multinomial Naïve Bayes (62%). The best performance of voting classifiers scored 73% using the classifiers SVM, RF and MNB.

Table 4 shows the results of the relation identification in subtask 2. Again SVM gave the best single classifier level performance.

## 5 Discussion

The total relation instances in the clean data and in the noisy data are almost the same (1228 and 1248, respectively). However, it is interesting to observe that the best performance in relation classification both at the single classifier level and in ensemble voting on noisy data (subtask 1.2) is significantly higher than on clean data (subtask 1.1). This behaviour is consistent also on the test data.

One explanation may be the differences in relation expressions in dataset $D_1$ and $D_2$. In the clean data ($D_1$), 25.66% of the entity mentions have three or more words with a maximum length of 13 words, whereas in the noisy data ($D_2$) only 0.96%

| Classifier | #Features | P | R | $F_1$ |
|---|---|---|---|---|
| SVM | 18,100 | 0.39 | 0.46 | **0.42** |
| DT | 5,800 | 0.50 | 0.29 | 0.36 |
| RF | 6,300 | 0.33 | 0.26 | 0.29 |
| MNB | 4,900 | 0.43 | 0.30 | 0.35 |
| KNN | 8,100 | 0.14 | 0.25 | 0.18 |
| Ensemble-all | – | 0.31 | 0.26 | 0.28 |
| Ensemble-best | – | 0.44 | 0.31 | **0.36** |

Table 4: Performance of positive class in relation identification on clean data (subtask 2). Ensemble-best is SVM+DT+MNB.

of the mentions have more than three words. The feature-based approach with n-grams as major feature source might not be able to capture the semantics of entity mentions having very large text spans. Furthermore, the context length between entity pairs in the clean data is larger than in the noisy data. Therefore, the shortest dependency paths and context n-grams—which are the two major feature sources—generate many insignificant features. Modeling the relation instances through a neural network could be a better alternative in this scenario.

Feature selection has a positive impact on prediction both in relation identification and in classification. SVM gave the best results at the single classifier level on all subtasks, but needs a larger feature space, whereas MNB performed reasonably although needing the smallest number of features for training the classifier.

## 6 Conclusion

In this work, we experimented with the relation identification and classification subtasks of SemEval 2018 Task 7 using a feature-based approach. A wide variety of features are explored, including lexical, syntactic, semantic, and other binary features. Two relation classification systems are developed on clean and noisy data and the third system is developed to identify relations in clean data. Five classifiers are trained for each subtask, with the final predictions made through voting based on the corresponding predictions of the individual classifiers. Experimental results shows that the lengths of the entity mentions and the lengths of the context in-between a pair of entities have significant impact on the relation identification and relation classification.

## References


Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. Semeval 2017 task 10: ScienceIE—extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada.

Biswanath Barik and Erwin Marsi. 2017. NTNU-2 at SemEval-2017 Task 10: Identifying synonym and hyponym relations among keyphrases in scientific documents. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 965–968, Vancouver, Canada.

Biswanath Barik, Erwin Marsi, and Pinar Öztürk. 2017. Extracting causal relations among complex events in natural science literature. In *International Conference on Applications of Natural Language to Information Systems*, pages 131–137, Liège, Belgium. Springer.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, Canada.

Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira, and Paula Carvalho. 2009. Relation detection between named entities: report of a shared task. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 129–137, Boulder, Colorado. ACL.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, Louisiana.

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *LREC 2016*, pages 3694–3701, Portorož, Slovenia.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99, Boulder, Colorado. ACL.

Jing Jiang and ChengXiang Zhai. 2007. A systematic exploration of the feature space for relation extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 113–120, Rochester, New York.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: Interactive Poster and Demonstration Sessions*, Barcelona, Spain. ACL. Paper 22.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. ACL.

Erwin Marsi, Pinar Øzturk, Elias Aamot, Gleb Sizov, and Murat Van Ardelan. 2014. Towards text mining in climate science: Extraction of quantitative variables and their relations. In *Proceedings of Bio-Text Mining*, Reykjavik, Iceland. European Language Resources Association.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1105–1116, Austin, Texas.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations*, pages 102–107, Jeju Island, Republic of Korea. ACL.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3(Feb):1083–1106.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland.