IS THERE TEXT IN WINE? – S+U LEARNING-BASED NAMED ENTITY RECOGNITION AND TRIPLET EXTRACTION FROM WINE AROMA DESCRIPTORS

Siddarth Venkateswaran[1], *Abdullah Al Foysal*[1], *Nazeer Basha Shaik*[2], *Ronald Böck*[1]

[1]*Genie Enterprise Inc., Branch Office Germany*
{venkat, afoysal, rboeck}@genie-enterprise.com,
[2]*nazeerbasha947@gmail.com*

Abstract: Wine making is usually considered a domain being far off the processing of speech and language. But in a particular aspect, the domains of speech processing and wine making are related, namely, in the description of wine aromas. These descriptors are used for creating wine expertise as well as more general (advertisement-like) textual representations. In the current paper, we use Natural Language Processing techniques, especially Named Entity Recognition, to identify Aspects and Opinions, reflecting wine characteristics. These are combined with analyses of respective relations (triplet extraction) building Aspect-Opinion-Pairs to establish indicative aroma descriptors, also trying to approach the complex interplay amongst these individual statements. In our experiments, we rely on the Falstaff corpus comprising a huge set of wine descriptions. This results in an average F1 score of around 0.85 for Aspect-Opinion classification. For triplet generation multiple strategies were compared, resulting in an average F1 score of 0.67 in this challenging task. For both tasks we rely only on a handful of manually annotated samples, applying pseudo-labeling methods from seed data to achieve automatic labeling.

1 Introduction

Wine making has a long history but is still a difficult and challenging business. For this, a multitude of expertise is necessary ranging from the perfect condition for harvesting the grapes to the process of fermentation to aging the wine in barrels, and finally the advertisement of the wine. The latter is related to a perfect description of the wine's aroma. Usually, winemakers and sommeliers exert their knowledge to achieve this, considering professional expertise as well as advertising texts, being intended for the general public [1].

As part of the project "PINOT"[1], we specifically regard the handling of wine aroma descriptors. The aim is the automatic processing of given descriptions to establish an automatically generated (synthesized) wine description based on multiple sensor inputs. In this sense, the entire process is divided into 1) the assignment of existing wine descriptions to scalar ratings, 2) the transfer of raw sensor inputs to scalar ratings, and 3) the automatic generation of novel wine descriptions from these sensor-based scalar ratings.

Our main objective is to evaluate how an automatic, natural language processing-based handling of wine aroma descriptors can be established and how the complex interplay of fine-graded opinion words like "deep dark" in relation to character-giving expressions like "chocolate brown" can be automatically linked (cf. Table 1). Therefore, in the current manuscript, we

[1]https://pinot-ai.com/ (last accessed 10 of January 2024).

DOI: 10.35096/othr/pub-7093, URN: urn:nbn:de:bvb:898-opus4-70931

focus on the pseudo-labeling of Aspect-Opinion-Pairs (AOPs), modeling the knowledge in wine aroma descriptors. Therefore, we relied on Named Entity Recognition (NER) models (e.g. [2]), Triplet Extraction Models (e.g. [3]), and the Simulated+Unsupervised Learning (S+U Learning) approach [4], applying those to the Falstaff corpus (cf. Section 3). Given this combination of methods, we were able to construct highly confident labels for AOPs on unlabeled data in an automatic fashion, reducing the manual effort drastically. As a result of the experiments, more than 80% of the data were pseudo-labeled with high confidence for the task of NER. It also yielded an increase in the F1 scores in the classification of Aspects and Opinions across each iteration. For the task of triplet extraction, the implementation of additional heuristics to the model predicted labels led to an increase in the F1 scores, and also a considerable amount of data being pseudo-labeled with high confidence. The subsequent sections of this paper focuses on the concept of S+U learning, outlining the steps involved in model training and discussing various strategies employed to enhance model performance.

2 Related Work

We focus on the main aspects contributing to the current manuscript and implementations, namely, NER, Triplet Extraction and pseudo-labeling.

NER is the task of fetching rigid designators from a given text description classified as person, location etc. [2]. Understanding and extracting structured information from unstructured text data is a crucial function of NER, which comes handy for many Natural Language Processing (NLP) applications such as question answering, text summarization, and machine translation [2]. [2] also provides a survey on the different deep learning approaches implemented for NER. Our task involves fetching Aspects and Opinions as entities within a wine descriptor, for which we have made use of the work of [5].

Triplet Extraction involves the identification of relations amongst different entities detected within a text [6]. A combination of NER and Triplet Extraction is used for knowledge extraction purposes in the textual content. [6] provides a survey on different approaches made for tasks related to Triplet Extraction. For our experiments, we made use of the implementation of [3] to find those Aspects and Opinions that were linked to one another.

Semi-Supervised Learning (SSL) involves training a neural network on a combination of manually labeled, and model-predicted highly-confident labeled data over multiple iterations until a termination condition is reached [7]. Pseudo-labeling is an SSL approach that has proven effective for training machine learning models with limited labeled data. Given our experience with pseudo-labeling in a different domain [8], a similar approach was taken to iteratively generate pseudo-labels for the task of NER and Triplet Extraction for texts within the wine domain.

3 Data Set

All experiments were implemented on the tastings data that were collected from Falstaff, an online accessible wine magazine. A total of 122,000 wine aroma descriptors were piled up, which were in German, comprising various regional influences in the wording. One of the key aspects was to collect wine aroma descriptors that were at sommelier level (cf. Table 1). From the entire corpus a subset of 5,000 samples was manually checked.

In the current research, we are aiming for AOPs. Therefore, the entities were classified as ‘Aspects’ which determined the unique characteristics of a wine, and ‘Opinions’ which further specify the Aspects. Based on these detected entities, we additionally aim to detect relevant triplets forming the AOPs. Consider from Table 1, for example, an aroma descriptor of type “Tiefdunkles Rubingranat, opaker Kern, violette Reflexe, zarte Randaufhellung”: the

Table 1 – Sample wine descriptors from Falstaff.

Original German Texts	English Translated Texts
Tiefdunkles Rubingranat, opaker Kern, violette Reflexe, zarte Randaufhellung.	Deep dark ruby-garnet colour, opaque core, violet reflections, delicate edge brightening.
Am Gaumen weich, rotbeerige Frucht, sehr lebendig, Orangenzestentouch im Finale.	Soft on the palate, red berry fruit, very lively, orange zest touch in the finish.

terms “Rubingranat”, “Kern”, “Reflexe” and “Randaufhellung” are Aspects, while “Tiefdunkles”, “opaker”, “violette” and “zarte” are characterizing Opinions. Hence, the AOPs formed within this descriptor are “Rubingranat - Tiefdunkles", “Kern - opaker", “Reflexe - violette" and “Randaufhellung - zarte". A sentence-level tokenization was performed on the collected corpus which yielded 360,000 sentences. As an initial seed, 200 sentences were picked from random positions and were manually annotated with the help of domain expertise.

4 Methods and Experimental Setup

4.1 Simulated+Unsupervised Learning

Originally, the idea of S+U Learning, introduced in [4], was developed in the context of the training process of Generative Adversarial Networks (GANs). Regarding the conceptual capabilities of those neural networks, GANs provide a framework to generate synthesized samples that can be used for training of neural approaches to improve their performance. For details on the entire original concept, we refer to [4] and [9]. In the current manuscript, we do not use a GAN approach, but adapted the S+U Learning method to our wine description framework.

In summary, given the S+U Learning approach, the (manual) effort of annotated data collections can be drastically reduced and in addition, “a nearly unlimited stream of training data could be produced with nearly zero marginal costs, given a sophisticated simulation” [9], which can be integrated into (theoretically) any learning paradigm (cf. e.g. [10]).

4.2 Network Approach

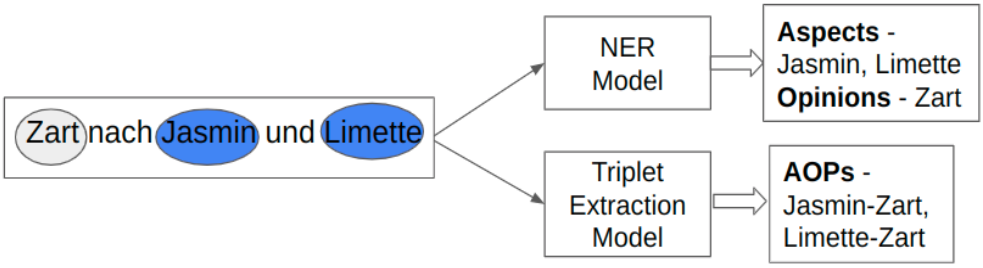
In line with the objectives of S+U Learning, simulated data was achieved using a combination of back-translation techniques (German-English-German) [11], fine-tuning an existing NER model [5] as well as an existing Triplet Extraction model [3], both with the help of German-BERT[2] embeddings. A separate NER model [5], and Triplet Extraction model [3] were trained iteratively on a combination of simulated and unlabeled data as a part of the pseudo-labeling steps.

Tasks related to NER were implemented by fine-tuning an existing Sequence Tagger model [5] using in particular the FLAIR framework [12]. These were fine-tuned using in particular the German-based BERT embeddings, with a mini-batch size of 4,096, and a mini-batch chunk size of 5 and with an early stopping technique for a patience of 5 epochs. The remaining hyperparameters were chosen in accordance with [5].

Additionally, tasks related to AOPs were implemented by fine-tuning the implementation of [3] with the help of German-based BERT embeddings. The remaining hyperparameters were retained as per the original implementations in [3].

[2]https://huggingface.co/bert-base-german-cased (last accessed 10 of January 2024).

Figure 1 – Workflow of the selected networks starting from individual samples. In the sentence, the Opinion (highlighted in grey) and Aspect (highlighted in blue) terms are marked.



4.3 Training the NER Models

NER models were trained across three iterations. During the first iteration, the model was adapted for 150 epochs purely on simulated data (training duration of 30 minutes). At the end of the first iteration, predictions were made on the unlabeled data. Those data which had an average confidence score greater than or equal to 0.95 were considered as confident pseudo-labeled data. During the second iteration, the model was fine-tuned on a combination of simulated data and confident pseudo-labeled data from the first iteration, for a total of 10 epochs. During the third iteration, the model was fine-tuned on a combination of simulated data, confident pseudo-labels from previous iterations, and also on a sample set of 200 low-confident samples that were picked and manually rectified at the end of the second iteration. The training period for the second and the third iterations lasted roughly 16 hours each.

For the evaluation of our results, we applied the commonly used F1 score, the harmonic ratio of recall and precision.

4.4 Training the AOP models

Different approaches were implemented to generate the pseudo-labels for the task of AOPs. This involved the naive approach, and the implementation of two additional strategies:

Naive Approach: This was similar to the approach taken in training the NER models. The first iteration involved training a model from scratch using only 200 simulated data. The second iteration involved fine-tuning this model using a combination of simulated data and confident pseudo-labeled data from the first iteration. The challenges faced in the outcomes of this approach (which will be discussed later) led to the implementation of two additional strategies.

Strategy 1: This involved the usage of 200 manually-rectified pseudo-labeled data at the end of iteration 1 in the naive approach. These comprised of 100 samples whose average confidence scores were in the range of 0.95 to 1.0 and were classified as high-confident after the application of heuristics (cf. Section 5.2), and 100 samples whose average confidence scores were in the range of 0.90 to 0.95. Therefore, the first strategy involved the usage of 400 data points - 200 each of simulated and rectified samples - for training a first iteration.

Strategy 2: The second strategy is based on a first training iteration purely relying on 200 rectified samples, attained from the first strategy, and evaluating the 200 manually labeled data. This was done to compare model performances that were trained on original data against simulated data. Additionally, heuristics were applied at the end of each iteration to narrow down the marking of high-confident pseudo-labels.

Figure 1 shows the workflow of the selected networks. Consider for example a descriptor of type “Zart nach Jasmin und Limette" (“delicate with jasmine and lime"): Here, the Aspects are “Jasmin" and “Limette", while the Opinion is “Zart". The triplets identified within this descriptor are “Jasmin - Zart", and “Limette - Zart".

As already applied in the NER experiments, we evaluated the performance using F1 scores.

Table 2 – Results on NER experiments considering respective F1 scores and a number of pseudo-labels.

	Iteration 1	**Iteration 2**	**Iteration 3**
F1 Score Aspect	0.891	0.894	0.896
F1 Score Opinion	0.796	0.809	0.811
# High Confident	101,058	302,164	330,244
# Low Confident	257,756	56,650	28,570
Training Time (h)	00:30	16:00	16:00

Table 3 – Results on Triplet Extraction experiments regarding F1 scores as well as the number of pseudo-labels.

Approach	**Naive Approach**		**Strategy 1**		**Strategy 2**	
	Iter 1	**Iter 2**	**Iter 1**	**Iter 2**	**Iter 1**	**Iter 2**
F1 Score AOP	0.92	0.34	0.63	0.63	0.21	0.67
# High Confident	3,184	4,077	369	383	3,849	4,358
# Medium Confident	441	57	32	14	258	287
# Low Confident	355,189	354,680	357,713	357,717	354,007	353,469

5 Results

In this section, we discuss the findings of our investigations and refer to Tables 2 and 3, highlighting the achievements.

5.1 Results on NER

For the NER experiments, there was an increase in the F1 scores across all entities from iterations 1 to 2. Considering the fact that the process of manual annotation is expensive and time-consuming, and also considering the limited availability of data for domain adaptation (i.e., the 200 manually annotated samples), already more than 100,000 data points were pseudo-labeled with high confidence by the end of the first iteration. Additionally, these pseudo-labeled data in combination with the simulated data led to more than 300,000 real data points being annotated by the end of iteration 2. Considering the time taken to manually annotate a handful of data (approximately 2 hours), and the time taken to train the models across two iterations, about 84% of the data were labeled with high confidence in a span of 18 hours. Based on the results of the second iteration, we ran a manual cross-checking, focusing on the lowest confident pseudo-labels, providing additional 200 samples reflecting those that caused difficulties. This resulted in a further adaptation of the model and further (slight) improvement in the F1 scores as can be seen in iteration 3 in Table 2. The real benefit of this approach lies in more than 330,000 sentences being pseudo-labeled with high-confidence, in addition to an increase in the F1 Score of classifying the Opinions detected within a text (cf. Table 2).

5.2 Results on Triplet Extraction

Drawbacks of the Naive Approach: Unlike with NER, pseudo-labeling AOPs using a naive approach proved to be a challenge. Training a model purely on simulated data yielded an F1 Score of 0.92 at the end of the first iteration (cf. Table 3). However, given the vast vocabulary present within the collected corpus, a back-translation technique to generate only 200 simulated data proved insufficient. Also, marking a sentence as high-confident purely on the basis of it achieving an average confidence score greater than 0.95 proved insufficient. This was evident

Table 4 – Example of AOPs prediction for an example sentence, at the end of iteration 1 using the naive approach.

Descriptor	Am Gaumen weich, rotbeerige Frucht, sehr lebendig, Orangenzestentouch im Finale		
Relations Formed	"Frucht - weich"	"Frucht - rotbeerige"	"lebendig - sehr"
Relation Scores	1.0	1.0	0.97
Correct Relations	No	Yes	Yes
Missed Relations	"weich - Am Gaumen", "Orangenzestentouch - im Finale"		

in a drastic drop in the F1 Score to 0.34 at the end of the second iteration (cf. Table 3). In the same time, we see an increase in the number of high confident samples, we are aiming for. For this, we analyzed the results per iteration in more details.

Additional investigations led to following considerations: Some challenges were observed in the way in which the predictions were made at the end of iteration 1, a sample is shown in Table 4. Here, it was observed that relations were formed at random positions within a text (for e.g. "Frucht - weich", cf. Table 4). Also, these randomly formed relations were predicted with high confidence scores. Another major challenge involved the inability of the model to detect all relations within the text (e.g., Missing Relations, cf. Table 4). The presence of a high number of such data being marked as high-confident pseudo-labeled data at the end of the first iteration (cf. Table 4) led to model poisoning in the second iteration (cf. Table 3). This observation led to the implementation of heuristics, which will be explained in the subsequent paragraph. In addition to setting a minimum confidence threshold score of 0.95, these heuristics added a level of strictness when it came to marking a data as high-confident pseudo-labeled data.

Implementation of Heuristics: We introduce a set of heuristics that were derived from the observations on the results at the end of the second iteration from the naive approach (cf. Table 3). These were done with an intention to alleviate the issues of randomly formed relations and missed relations while marking a data as high-confident pseudo-labeled. Also, they were deliberately implemented to select those data points that had a high number of relations formed within them.

The suggested and implemented heuristics comprise as follows:

- Check if relations are not formed at random points within a text.
- Check for each detected entity, if it is paired at least with one other entity.
- Check if the detected entity is a complete word.
- Check if at least 50% of all words within text is detected as entities.
- Check if number of relations formed within a text, $N_{\text{Relations}}$ is greater than or equal to the highest detected entity within a text such that:

$$N_{\text{Relations}} \geq \max(N_{\text{Aspects}}, N_{\text{Opinions}}),$$

  where N_{Aspects} and N_{Opinions} are the number of Aspects and Opinions detected in a text.

Observations with Strategies 1 and 2: In strategies 1 and 2, we applied heuristics which improved the abilities of the triplet generation. Despite low F1 scores, there was an improvement over the baseline for each strategy at the end of the second iteration (cf. Table 3). More importantly, these led to a consistent model performance across iterations in strategy 1 and an

improvement in strategy 2 (cf. Table 3). Also, in strategy 2, it was observed that running these experiments for 2 iterations looked optimum, for a third iteration led to a further drop in the model performance.

Regarding the different strategies and results, we can argue as follows: A naive approach being transferred from the NER experiments is not suitable, given the wine descriptions in the Falstaff corpus (cf. Section 3). We need heuristics to improve the quality of triplets in the current task.

So, let us consider the two remaining strategies. The main difference is in the handling of synthetic data. Strategy 1, in which a number of confident pseudo-labeled data that were generated at the end of the first iteration, led, however, to a more consistent model. From our perspective, this is related to the approach of staying comfortable within the bubble of S+U Learning synthesized samples. For the matter of a mere increase of data, the approach produces reasonable results (cf. Table 3). However, in order to go beyond the bubble, there is a benefit from the addition of real, fresh data to the already generated pseudo-labeled data in combination with the application of heuristics. This yielded an improvement in both, the number of high-quality pseudo-labeled data being generated as well as in the performance of the models across iterations (cf. Table 3). Given these results, we expect that combined handling of real and pseudo-labeled data, as in strategy 2, might be the most beneficial approach for an automatic generation of AOPs.

6 Conclusion

The task of fetching AOPs from wine descriptors helps in knowledge extraction on the aromatic characteristics of a wine. This paper aimed to achieve this with the help of S+U Learning as a two-step process (each step had multiple iterations): first by making use of an NER model to detect all the named entities within a given wine descriptor in the form of Aspects and Opinions, and then by making use of a Triplet Extraction method by forming pairs between Aspects and their corresponding Opinions, forming AOPs (cf. Figure 1). While the NER experiments yielded a slight increase in the F1 scores, we achieved, in contrast, a high number of highly confident pseudo-labeled data with very low (manual) effort (cf. Table 2). While the Triplet Extraction experiments proved to be challenging due to the complex nature of wine aroma descriptions, the addition of a small amount of fresh, real samples did yield an improvement in the model performance across iterations (cf. Table 3). This was achieved by combining automatic training approaches and implementing additional heuristics, leading to improved quality of the pseudo-labels that were generated at the end of each iteration.

As a next step, with the availability of a good number of high-confident triplets, the aim is to train a model for a joint task of NER and AOPs prediction in a given wine descriptor.

7 Acknowledgements

We acknowledge support by the PINOT project funded by the German Federal Ministry of Food and Agriculture (BMEL) under grant number 28DK107C20.

References

[1] HONORÉ-CHEDOZEAU, C., M. DESMAS, J. BALLESTER, W. V. PARR, and S. CHOLLET: *Representation of wine and beer: influence of expertise*. *Current Opinion in Food Science*, 27, pp. 104–114, 2019. doi:https://doi.org/10.1016/j.cofs.2019.07.002. URL

https://www.sciencedirect.com/science/article/pii/S2214799319300116. Sensory Science and Consumer Perception • Food Physics materials Science.

[2] LI, J., A. SUN, J. HAN, and C. LI: *A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering*, 34(1), pp. 50–70, 2020.

[3] XU, L., Y. K. CHIA, and L. BING: *Learning span-level interactions for aspect sentiment triplet extraction*. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4755–4766. Association for Computational Linguistics, Online, 2021.

[4] SHRIVASTAVA, A., T. PFISTER, O. TUZEL, J. SUSSKIND, W. WANG, and R. WEBB: *Learning from simulated and unsupervised images through adversarial training*. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, p. s.p. Honolulu, USA, 2017.

[5] SCHWETER, S. and A. AKBIK: *Flert: Document-level features for named entity recognition. arXiv preprint arXiv:2011.06993*, 2020.

[6] NAYAK, T., N. MAJUMDER, P. GOYAL, and S. PORIA: *Deep neural approaches to relation triplets extraction: A comprehensive survey*. *Cognitive Computation*, 13, pp. 1215–1232, 2021.

[7] KINGMA, D., S. MOHAMED, D. JIMENEZ REZENDE, and M. WELLING: *Semi-supervised learning with deep generative models*. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. D. LAWRENCE, and K. Q. WEINBERGER (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3581–3589. Curran Associates, Inc., 2014.

[8] VENKATESWARAN, S., R. BÖCK, T. KESSLER, and O. KRINI: *Pseudo-labelling and transfer learning based speech emotion recognition*. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pp. 19–26. TUDpress, Dresden, 2021.

[9] KROKOTSCH, T. and R. BÖCK: *Generative adversarial networks and simulated+unsupervised learning in affect recognition from speech*. In *Proc. of the 8th International Conference on Affective Computing & Intelligent Interaction*, p. s.p. IEEE, Cambridge, United Kingdom, 2019.

[10] LEE, K., H. KIM, and C. SUH: *Simulated+unsupervised learning with adaptive data generation and bidirectional mappings*. In *International Conference on Learning Representations*, pp. 1–15. ICLR, 2018.

[11] HOANG, C. D. V., P. KOEHN, G. HAFFARI, and T. COHN: *Iterative back-translation for neural machine translation*. In *2nd Workshop on Neural Machine Translation and Generation*, pp. 18–24. Association for Computational Linguistics, 2018.

[12] AKBIK, A., T. BERGMANN, D. BLYTHE, K. RASUL, S. SCHWETER, and R. VOLLGRAF: *FLAIR: An easy-to-use framework for state-of-the-art NLP*. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59. 2019.