Semantic Frame Extraction in Multilingual Olfactory Events

Stefano Menini
Fondazione Bruno Kessler
Trento, Italy
menini@fbk.eu

Abstract

In this work we present a system for multilingual olfactory information extraction covering six European languages, introducing new models to extract olfactory information from large amounts of text in a structured and scalable way. For the task, we rely on a supervised multi-task approach to detect olfactory-related text adopting a FrameNet-like structure, so that both the lexical units triggering the smell event and a related set of frame elements are identified.

Keywords: olfactory events, semantic frame extraction, multi-task

1. Introduction

In recent years, there has been a growing interest in developing resources specifically designed to capture sensory aspects of the language (Winter, 2019). Among the five senses, olfaction is, together with taste, the sense having less specific vocabulary to describe it in Western languages (Majid and Burenhult, 2014). Furthermore, it is an extremely interesting domain to explore due to the ephemeral nature of smells and the role they played in signalling identity, community and otherness in the past (Tullett et al., 2022).

Despite the interest in studying this domain, little effort has been devoted to develop tools and models that can extract olfactory information from large amounts of text in a structured and scalable way. In this work, we therefore focus on this task by presenting a supervised system for multilingual olfactory information extraction covering six European languages, namely *English*, *French*, *Italian*, *Dutch*, *German* and *Slovene*.

The system detects the parts of text involved in an olfactory event adopting a FrameNet-like structure: it identifies the lexical units triggering the smell event and a set of semantic roles associated to the olfactory event that have been previously defined by domain experts (e.g. the *smell source* or the *effect* provoked by the smell). To this end, a transformer-based model is fine-tuned on an existing benchmark with olfactory information by adopting a multi-task framework. The ability to extract and analyze semantic frames related to olfactory events represents a step forward towards enhancing our understanding of the olfactory world through quantitative analysis of large collections of texts.

The models presented in this paper are available at this link: https://zenodo.org/records/10598306.

2. Related Work

Two tasks are relevant to our work: event extraction and frame-semantic parsing.

The event extraction task consists in the identification of event mentions in text (Ahn, 2006; Liao and Grishman, 2010; Nguyen and Nguyen, 2019; Lu et al., 2021). But while most works in this field focus on determining the event type across different domains such as social media (de Bruijn et al., 2019) or history (Lai et al., 2021), here we focus on the detection of the single "Olfactory Event", encompassing all the possible shapes of olfactory experiences and not figuring in previous resources as FrameNet (Ruppenhofer et al., 2016).

The second task, Frame Parsing, is about automatically recognizing the presence in texts of semantic frames, conceptual structures that provide a framework to describe prototypical situations and the specific roles involved. Examples are Das et al. (2014) where the task is approached in two-stages, first identyfing lexical targets and then predicting frame-semantic structures or Swayamdipta et al. (2018) that incorporate syntactic information into the task.

3. Training Data

The data we used to train the models in this paper is the multilingual olfactory benchmark from Menini et al. (2022). The dataset contains annotations of olfactory events and situations in texts from ten different domains (e.g. narrative, medicine or travel) in 6 European languages, namely English, Italian, French, German, Dutch and Slovene. Olfactory annotation follows a FrameNet-like approach (Ruppenhofer et al., 2016)[1], focusing on the semantic roles involved in the olfactory situations.

As in FrameNet, *frames* are used as synonyms for schemata or scenarios. A frame includes two

[1] https://framenet.icsi.berkeley.edu

LREC-COLING 2024, pages 14622–14627

main components:

Lexical Units (LUs): words, multiwords or idiomatic expressions that evoke a specific frame, in this case an olfactory frame (e.g. '*smell*', '*odour*', '*perfume*'). Also defined *"smell words"* by the authors of the benchmark.

Frame Elements (FEs): frame-specific semantic roles related to the olfactory frames. An overview of the frame elements included in the dataset is presented in Table 1.

This benchmark, despite containing only around 1,700 olfactory events per language, can be used to train a supervised system aimed at recognising the information that was originally labeled by human annotators. The distribution of the Frame Elements in the dataset is not balanced with '*Smell Source*' and '*Quality*' being the most frequent ones followed by the other 7 Frame Elements that are significantly less represented. This is true for all the six languages taken into consideration, showing how smell sources and qualities can be considered the core elements for the given olfactory-related lexical units. On average, each lexical unit is associated with between 2.1 and 2.7 frame elements.

In the next section we describe the system implemented to classify the lexical unit and all the 9 labels presented in Table 1. Differently from the work presented in Menini et al. (2023), the focus will be not only on the more represented core semantic roles for olfactory situations, i.e. *smell words* (the lexical unit), *smell sources* and *qualities*, but also the other frame elements that are less frequent in the benchmark, to investigate whether the system is able to identify them even if few training instances are available.

4. Model Training

To extract olfactory information (lexical units and frame elements) from text we compare two different paradigms.

- In the first setting, we consider lexical units detection and frame element classification as part of the same multiclass token classification task (*single task* approach).
- In the second one, instead, we adopt a *multi-task* approach, considering the classification of lexical units and of each frame element as separate tasks.

Given the advantages and the good performance obtained with pre-trained language models (LM) based on the Transformer architecture in several downstream NLP tasks (Vaswani et al., 2017), we use in both settings a transformer-based model by fine-tuning it to perform a token classification task.

We experiment both with monolingual and multilingual variants of either BERT or RoBERTa, depending on their availability for each language. The single task and multi-task approaches are therefore tested in two configurations. In the first one, the model for each language is obtained by fine-tuning on monolingual data with monolingual models, while in the second configuration the fine-tuning is done on the olfactory benchmarks of all the six languages together using a multilingual model that is then tested on each language separately. The models used for each language are:

En: bert-base-cased[2] (Devlin et al., 2019)

It: bert-base-italian-cased[3] (Schweter, 2020)

Nl: bert-base-dutch-cased[4] (de Vries et al., 2019)

Fr: flaubert_base_cased[5] (Le et al., 2020)

Sl: sloberta[6] (Ulčar and Robnik-Šikonja, 2021)

De: bert-base-german-cased[7] (Chan et al., 2020)

Multilingual: bert-base-multilingual-cased (mBert)[8] (Devlin et al., 2019)

The two classification frameworks are evaluated using the same 10-fold configuration and sharing training/validation/test splits, so that results are comparable. Each data split has 80% of the lexical units and related frame elements (FE) used as training data, 10% for validation and 10% as test. The splits are not completely random as we sought to keep the same FE distribution in every run. The same splits are kept for all the tested configurations.

Similar to tasks such as named-entity recognition, where we assign a label to each token but at the same time we need to define where the span of an entity starts and ends, the two classification approaches share the same IOB labeling data format, in which tokens in a span are marked with Inside–Outside–Beginning of Olfactory frame element labels.

4.1. Single Task Classification

The first set of models for olfactory information extraction has been designed as a token classification

[2] https://huggingface.co/bert-base-cased

[3] https://huggingface.co/dbmdz/bert-base-italian-cased

[4] https://huggingface.co/GroNLP/bert-base-dutch-cased

[5] https://huggingface.co/flaubert/flaubert_base_cased

[6] https://huggingface.co/EMBEDDIA/sloberta

[7] https://huggingface.co/bert-base-german-cased

[8] https://huggingface.co/bert-base-multilingual-cased

Frame Element	Example Sentence
Smell Source	The person, object or place that has a specific smell. *The odour [of tar] and [pitch] was so strong.*
Odour Carrier	The carrier of an odour, either an object or atmospheric elements. *The unpleasant smell [of the vapour] of linseed oil extended for a considerable distance.*
Quality	A quality associated with a smell and used to describe it. *Earth has a [strong], [aromatic] odour.*
Perceiver	The being that perceives an odour, who has a perceptual experience, not necessarily. *The scent is described by [Dr. Muller] as delicious.*
Evoked Odorant	The object, place or similar that is evoked by the odour, even if it is not in the scene. *In offensive perspiration of the feet [a peculiar cabbage-like] stench is given off.*
Location	The location where the smell event takes place. *And, particularly, [at the foot of the garden], where he felt a so very offensive smell.*
Time	An expression describing when the smelling event occurred. *Galeopsis smells fetid [at first handling], [afterwards] aromatic.*
Circumstances	The state of the world under which the smell event takes place. *[When stale] the lobster has a rank stench.*
Effect	An effect or reaction caused by the smell. *An ill smell [gives a nauseousness].*

Table 1: Overview of the Frame Elements (FEs) related to Olfactory situations and events with corresponding examples. Lexical units are underlined and the FE of interest is in square brackets. The same definitions hold for all languages included in the benchmark. For more details on FEs descriptions see (Tonelli and Menini, 2021).

	NL	EN	FR	DE	IT	SL
Smell Word	1,788	1,530	845	2,659	1,254	1,973
Smell Source	1,922	1,313	710	2,297	952	1,638
Quality	1,071	1,084	450	1,730	707	936
Perceiver	336	362	140	399	153	266
Circ.	399	248	88	274	202	228
Odor Carrier	351	310	106	170	195	408
Effect	243	187	53	425	104	214
Evoked Odorant	228	91	103	258	74	285
Place	255	302	172	200	158	394
Time	127	126	49	131	119	75

Table 2: Overview of annotated instances of lexical units (*Smell Words*) and frame elements in the benchmark for each language.

task, where the system has to assign to each token in the text one out of 21 labels, i.e. 20 being either *"begin"* or *"inside"* of each lexical unit and frame element, plus the *"outside"* label. In fact, we can define the task as a single task multiclass classification problem.

Each model has been fine-tuned with a token classification head on top.[9] During training, a hyperparameter search was applied to the first fold of each language with the model under investigation over the search space: learning rate $[1e-5, 2e-5, 3e-5, 4e-5, 5e-5]$, batch size $[8, 16, 32]$, training epochs up to 20. Warmup for 10% of the training steps was applied. After determining the hyperparameters for each model, it was fine-tuned 10 times, each time with a different data fold, and average scores were computed.

4.2. Multi-task Classification

The second configuration we test is multi-task learning (Caruana, 1993, 1997). We train a neural network to learn different tasks in parallel while using a shared representation, so that each task updates the model's shared parameters with respect to every task, ideally leading to a more robust representation with less over-fitting. In this configuration, each task corresponds to the classification of a single olfactory element, namely *Smell Word*, *Smell Source*, *Quality*, *Odour Carrier*, *Evoked Odorant*, *Location*, *Perceiver*, *Time*, *Circumstances*, *Effect*.

Frame elements related to the olfactory domain can be ambiguous, with most of the smell sources not being olfactory-specific items and with qualities being either generic such as "pleasant" or borrowed from other senses as "sweet". We adopt a multitask approach with the hypothesis that simpler tasks, i.e. detecting the lexical units (*smell words*), can act as auxiliary task and share information for the classification of more difficult and ambiguous frame elements. Indeed, while lexical units are usually expressed by single terms, frame elements typically match text spans, usually corresponding to

[9]The Huggingface Transformers library was used to implement the token classification task. https://huggingface.co/docs/transformers/tasks/token_classification

Lan.	App.	Train Data	Smell Word	Smell Source	Quality	Odour Carrier	Evoked Odorant	Loc.	Perc.	Time	Circ.	Effect
EN	MT	mono	0.871	0.571	0.758	**0.482**	**0.572**	0.542	**0.510**	0.434	0.461	**0.405**
		multi	0.865	**0.574**	**0.759**	0.462	0.517	**0.546**	0.488	**0.528**	**0.480**	0.339
	ST	mono	0.867	0.525	0.703	0.392	0.293	0.368	0.410	0.304	0.266	0.140
		multi	**0.881**	0.530	0.698	0.392	0.359	0.410	0.390	0.309	0.261	0.138
IT	MT	mono	0.871	0.559	0.800	0.343	0.564	**0.439**	0.241	0.613	0.259	**0.246**
		multi	**0.887**	**0.575**	**0.801**	**0.382**	**0.625**	0.304	0.240	**0.642**	**0.309**	0.201
	ST	mono	0.854	0.387	0.739	0.249	0.383	0.193	**0.254**	0.461	0.148	0.141
		multi	0.880	0.407	0.755	0.269	0.299	0.186	0.231	0.398	0.201	0.149
FR	MT	mono	**0.839**	0.459	0.567	**0.440**	**0.536**	**0.373**	**0.380**	0.481	0.279	0.235
		multi	0.838	**0.472**	**0.591**	0.379	0.505	0.322	0.258	**0.561**	**0.306**	**0.273**
	ST	mono	0.734	0.314	0.336	0.327	0.414	0.302	0.291	0.384	0.118	0.142
		multi	0.820	0.417	0.488	0.352	0.367	0.251	0.268	0.336	0.111	0.150
NL	MT	mono	0.788	0.376	0.632	0.191	0.444	**0.238**	0.303	0.313	0.133	0.149
		multi	**0.789**	**0.407**	**0.638**	**0.214**	**0.468**	0.236	**0.308**	**0.342**	**0.154**	**0.192**
	ST	mono	0.725	0.225	0.545	0.041	0.091	0.068	0.104	0.096	0.053	0.045
		multi	0.765	0.235	0.556	0.072	0.063	0.082	0.064	0.071	0.030	0.039
DE	MT	mono	0.812	0.454	0.668	0.157	0.454	**0.308**	**0.358**	0.184	**0.300**	0.241
		multi	**0.814**	**0.470**	**0.677**	**0.215**	**0.490**	0.293	0.351	**0.255**	0.273	**0.253**
	ST	mono	0.778	0.273	0.479	0.186	0.150	0.092	0.164	0.162	0.040	0.036
		multi	0.797	0.268	0.443	0.141	0.086	0.092	0.133	0.095	0.031	0.030
SL	MT	mono	**0.707**	**0.501**	**0.525**	**0.320**	**0.506**	**0.401**	**0.355**	**0.280**	**0.153**	**0.151**
		multi	0.695	0.442	0.491	0.273	0.445	0.368	0.245	0.214	0.103	0.132
	ST	mono	0.675	0.406	0.451	0.119	0.277	0.236	0.170	0.155	0.068	0.074
		multi	0.655	0.358	0.448	0.186	0.263	0.212	0.195	0.137	0.051	0.086

Table 3: Results (F1) of the classifiers on the lexical unit and 9 frame elements for both Single Task (ST) and Multi-task (MT) approaches. Each result is the average of 10 different runs done on 10 different data splits). *mono* = monolingual data and model; *multi* = multilingual training data and model

constituents. Therefore, not only label identification but also span detection make the task of frame element classification complex.

To fine-tune the models, we use MaChAmp (van der Goot et al., 2021), a toolkit for fine-tuning in multi-task settings, and the classification of each frame element was again configured as IOB task. MaChAmp can be configured with a different loss weight parameter for each task to define the main/auxiliary tasks. For each task, we compare two different values of loss weight: 1 and 0.75, testing different combinations over the 10 tasks. A hyperparameter search was applied to one of the splits with the following search space: learning rate $[1e{-}3, 1e{-}4, 1e{-}5]$, batch size $[16, 32]$ and number of training epochs $range(1, 30)$. All configurations reported in Table 3 use a learning rate of $1e-4$ and a batch size of 32, and all the loss weight set to 1, which yield the best performance.

5. Results

The result of the different configurations are reported in Table 3. As expected, *smell words* and the more represented *smell sources* and *qualities* are better classified than other Frame Elements with less training instances. After doing a manual check of the mistakes in the output, we notice that a large portion of the errors is not due to missing frame elements (or erroneously identified when not present) but rather to mismatches in the boundaries of the FE spans, e.g. predicting as smell source *"flowers"* rather than *"of flowers"*. This type of errors has a larger impact on FE such as *"circumstances"* and *"effect"*, often consisting of longer portions of text resulting in more incorrectly classified tokens.

Another aspect emerging from the results is that in all the languages the multi-task classifier is more effective than a single task classifier, supporting the idea that treating the classification of each FE as a separate task is beneficial because each FE encodes very peculiar information. Finally, fine-tuning the model on multiple languages has been proven helpful only on Italian, German and Dutch, while other languages obtain the best result with their respective monolingual models.

6. Conclusions

In this work we present the first system for multilingual olfactory information extraction. To our knowledge, this is the first time that frame-like annotation is tackled through multi-task classification. In the future, it would be interesting to check whether this approach is beneficial also when applied to the full FrameNet annotation. We also plan to use the system to perform large-scale studies on olfactory language, comparing perceptions across different cultures.

7. Acknowledgements

This research has been supported by the European Union's Horizon 2020 program project ODEUROPA under grant agreement number 101004469.

8. Bibliographical References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8.

R Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Citeseer.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Jens A de Bruijn, Hans de Moel, Brenden Jongman, Marleen C de Ruiter, Jurjen Wagemaker, and Jeroen CJH Aerts. 2019. A global database of historic and real-time flood events based on social media. *Scientific data*, 6(1):311.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 789–797.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*.

Asifa Majid and Niclas Burenhult. 2014. Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2):266–270.

Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroğlu, and Sara Tonelli. 2023. Scent mining: Extracting olfactory events, smell sources and qualities. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 135–140.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6851–6858.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute.

Stefan Schweter. 2020. Italian bert and electra models.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A Smith. 2018. Syntactic scaffolds for semantic structures. *arXiv preprint arXiv:1808.10485*.

Sara Tonelli and Stefano Menini. 2021. FrameNet-like annotation of olfactory information in texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–20, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

William Tullett, Inger Leemans, Hsuan Hsu, Stephanie Weismann, Cecilia Bembibre, Melanie A. Kiechle, Duane Jethro, Anna Chen, Xuelei Huang, Jorge Otero-Pailos, and Mark Bradley. 2022. Smell, history, and heritage. *American historical review*, 127(1):261–309.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Sloberta: Slovene monolingual large pretrained masked language model.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bodo Winter. 2019. *Sensory linguistics: Language, perception and metaphor*, volume 20. John Benjamins Publishing Company.

9. Language Resource References

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.