Scent Mining: Extracting Olfactory Events, Smell Sources and Qualities

Stefano Menini[†], **Teresa Paccosi**[†‡], **Serra Sinem Tekiroğlu**[†] **Sara Tonelli**[†]
[†]Fondazione Bruno Kessler, Trento, Italy,
[‡]University of Trento, Italy
{menini,tpaccosi,tekiroglu,satonelli}@fbk.eu

Abstract

Olfaction is a rather understudied sense compared to the other human senses. In NLP, however, there have been recent attempts to develop taxonomies and benchmarks specifically designed to capture smell-related information. In this work, we further extend this research line by presenting a supervised system for olfactory information extraction in English. We cast this problem as a token classification task and build a system that identifies smell words, smell sources and qualities. The classifier is then applied to a set of English historical corpora, covering different domains and written in a time period between the 15th and the 20th Century. A qualitative analysis of the extracted data shows that they can be used to infer interesting information about smelly items such as tea and tobacco from a diachronical perspective, supporting historical investigation with corpus-based evidence.

1 Introduction

In recent years, research on sensory-related analysis of texts has become more and more relevant within the NLP community. Indeed, studies in linguistics, primarily aimed at assessing how sensory language differs across languages and how the different senses are described and compared (Majid and Burenhult, 2014; Strik Lievers and Winter, 2018; Winter et al., 2018; Winter, 2019), have then paved the way for more computationally-oriented analyses, aimed for example at structuring the sensory vocabulary in machine-readable taxonomies (Tekiroğlu et al., 2014a,b; McGregor and McGillivray, 2018; Menini et al., 2022a), using distributional semantics to explore sensory blending (Girju and Lambert, 2021), or extracting sensory information about cities using data from social media (Quercia et al., 2015).

Several works have dealt with olfaction, which is a sense that is traditionally less represented in the vocabulary of Western European languages and in texts (Winter et al., 2018), making it a very interesting domain to investigate with computational means. The first work trying to capture smelly experiences using two semi-supervised approaches was presented in Brate et al. (2020), while annotation guidelines and a multilingual benchmark for olfactory information have been recently released (Tonelli and Menini, 2021; Menini et al., 2022b). Along this research line, we present a supervised system for olfactory information extraction in English trained on the above benchmark. We cast this task as a token-classification problem, labelling smell words that evoke olfactory events and two related semantic roles: smell sources and qualities. We present not only a standard evaluation by measuring F1 on each element, but also a qualitative analysis, by applying our system to four historical corpora in English and manually interpreting the extracted information.

The code and the model used to extract olfactory information from texts are available at https://github.com/dhfbk/scent-mining.

2 Dataset

In order to train a system for olfactory information extraction, we use the English benchmark presented in Menini et al. (2022b).[1] The benchmark contains 85 documents, distributed evenly over a time period between 1620 and 1920 and covering 10 domains: *Household & Recipes*, *Law*, *Literature*, *Medicine & Botany*, *Perfumes & Fashion*, *Public health*, *Religion*, *Science & Philosophy*, *Theatre*, *Travel & Ethnography*.

The benchmark was annotated with olfactory information following the guidelines presented in Tonelli and Menini (2021). This scheme is inspired by frame semantics (Fillmore and Baker, 2001) and the FrameNet annotation project (Ruppenhofer

[1]Available at https://github.com/Odeuropa/benchmarks_and_corpora

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 135–140

et al., 2006),[2] whose goal is to capture situations and events present in texts. In the benchmark that we use for our experiments, only one event type was considered, i.e. *Olfactory event*, which according to the guidelines can be evoked by a *smell word*, or Lexical Unit (LU). Such smell word may be connected to one or more *Semantic roles* (so-called Frame Elements, FEs) participating in such event. English smell words include nouns such as 'stink', 'odour','stench','whiff', verbs such as 'to smell','to reek','to sniff', adjectives such as 'scented','odorous','reeking', and adverbs such as 'pungently'.

The benchmark contains 1,530 olfactory events. Concerning semantic roles, the annotation scheme foresees nine of them, namely *Smell source*, *Quality*, *Evoked odorant*, *Odour carrier*, *Perceiver*, *Time*, *Location*, *Effect* and *Circumstances*. The most frequent ones are *Smell source* and *Quality*, which are both represented by respectively 1,313 and 1,084 instances in the benchmark, while all the others are much more sparse. For this reason, we include in our first system for olfactory information extraction only the recognition of smell words, *Smell source* and *Quality*, leaving the other roles to future extensions.

We report in Table 1 the definition of these two FEs. According to these guidelines, if we consider the sentence below, we would annotate '[The coffee]' as *Smell source* and '[pungent]' as *Quality*, while 'smell' would be the lexical unit evoking the olfactory event.

[The coffee] had a [pungent] smell.

Frame Element	Definition and Example
Smell Source	The person, object or place that has a specific smell. It can also refer to (non)human/object that produces an odour (e.g. plant, animal, perfume, human). The entity or phenomenon that the perceiver experiences through his or her senses.
Quality	A quality associated with a smell and used to describe it. This is typically expressed by qualitative adjectives and it is often preceded by an intensifier such as 'very, really'. Qualities include intensity ('weak', 'distinct'), volume/reach ('far reaching'), duration ('lasting', 'permanent'), state ('old', 'deteriorated'), character ('dry', 'garlicky'), hedonic characteristics ('malodorous', 'aromatic').

Table 1: Definition of Smell Source and Quality from the benchmark annotation guidelines.

3 System for Olfactory Information Extraction

The model for olfactory information extraction has been designed as a token classification task, i.e. a natural language understanding task in which a label is assigned to each token in a given text. While past works in semantic frame parsing usually treated lexical unit detection and frame element annotation as two separate tasks (Das et al., 2014), we consider them both at the same level and build a single classification model. We use the IOB labeling data format, in which tokens in a span are marked with Inside–Outside–Beginning of smell-related elements. The model labels each token as O (outside), B-FRAME_ELEMENT (beginning of a span of an olfactory element) or I-FRAME_ELEMENT (inside of a span of an olfactory element) given an input sentence. As introduced above, we label both smell words and the two most frequent frame elements, namely *Smell Source* and *Quality*.

Considering the advantages of pre-trained language models (LM) based on the Transformer architecture for downstream NLP tasks (Vaswani et al., 2017), we use the pre-trained BERT models (Devlin et al., 2019) in our experiments. Each model has been fine-tuned with a token classification head on top.[3] We experiment both with a monolingual language model (bert-base-uncased)[4] and its multilingual variant (bert-base-multilingual-uncased)[5] and fine-tune these models for the token classification task.

We perform five-fold cross-validation, using 80% of the data for training, 10% for validation and 10% for testing. During training, a hyperparameter search is applied to Fold-0 with the model under investigation over the search space: learning rate $[1e-5, 2e-5, 3e-5, 4e-5, 5e-5]$, batch size $[4, 8]$, number of training epochs $range(1, 10)$. Warmup for 10% of the training steps was applied. After determining the hyperparameters for each model, it is fine-tuned 5 times, each time with a different data fold, and average scores are computed.

Table 2 shows the classification results obtained

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

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

[4] https://huggingface.co/bert-base-uncased

[5] https://huggingface.co/bert-base-multilingual-uncased

	F1			
	Smell Word	Source	Quality	Overall
BERT	0.877	**0.503**	**0.686**	**0.689**
mBERT	**0.885**	0.490	0.672	0.682

Table 2: Classification results (macro F1) on English. We distinguish between using monolingual BERT and mBERT.

on *Smell Words*, *Smell Sources* and *Qualities*, as well as the overall score obtained by averaging the system performance on these three elements. Our evaluation is based on "exact match", i.e. the smell words and the other roles are considered correctly identified only if they match completely with the annotation in the gold standard. If there is a partial overlap of the tokens, the labelling is considered not correct.

We compare the results obtained with monolingual BERT and its multilingual version (mBERT). Note that performance on smell words is better than on the other frame elements because the former are mostly single words, while *Smell Source* and *Quality* are typically expressed by phrases and also the identification of the correct span can be very challenging. Overall, there is only a slight difference between BERT and mBERT, with BERT performing better. Therefore, we adopt this model for our next analysis.

4 Olfactory Information Extraction

We launch the BERT-based model on a set of historical corpora of English. Our goal is to analyse the smell-related information extracted by our system and to perform some qualitative study of the results. We focus on four freely available corpora:

Project Gutenberg:[6] A volunteer effort to digitize and archive cultural works, it contains different repositories, mainly in the literary domain (4,943 books, 366M tokens).

The Royal Society Corpus:[7] A repository of scientific periodicals issued between 1665 and 1869 (9,782 documents for a total of 31M tokens);

A pre-processed subset of *the Old Bailey Papers* dataset,[8] containing the court proceedings published between 1720 and 1913 (638 books 3.1M tokens).

Early English Books Online (EEBO),[9] containing documents published between 1475 and 1700 in different domains such as literature, philosophy, politics, religion, geography, history, politics, mathematics (60,329 documents for a total of 1.4B tokens)

In Table 3 we provide an overview of the olfactory information extracted from the above set of corpora. The data are divided into two groups based on their publication date, which will be used for the analysis presented in Section 5:

	1500-1799	1800-1930	Total
Smell Sentences	91,018	32,442	123,460
Smell Sources	66,070	27,776	93,846
Qualities	49,275	19,039	68,314

Table 3: Sentences containing at least a smell word and the number of associated *Smell Sources* and *Qualities*.

5 Case study: Perception shift

Inspired by past approaches to semantic shift detection, we examine potential changes in the way a specific smell source is described in texts. We argue that these variations may reflect a shift in the perception of specific smells, as already highlighted in historical research using qualitative approaches (Tullett, 2019b).

In our analysis we compare the meaning of the smell sources before and after 1800. We select this period because it represents a significant turning point in the cultural attitudes towards scent, especially in England. The sense of smell acquired an increasingly social significance and played a role in shaping both individual identities and those of specific places (Tullett, 2019a). For this purpose we split the extracted data in two parts, the first one covering the period from 1500 to 1799 and the second one from 1800 to 1930.

To identify perception shifts in the olfactory information extracted from our data, we follow the work by El-Ebshihy et al. (2018) on semantic shifts. First, we reduce the vocabulary of the text extracted by lemmatizing it with Stanza (Qi et al., 2020). Then, for each time period, we create an embedding space with FastText, using the skip-gram model and an embedding size of 100 (Bojanowski et al., 2016). To be able to compare the embeddings from the two time periods, we align the 1800-1930 space

[6] https://www.gutenberg.org/

[7] http://fedora.clarin-d.uni-saarland.de/rsc_v4/

[8] http://fedora.clarin-d.uni-saarland.de/oldbailey/downloads.html

[9] https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/

Smell Source	Cosine Similarity	Smell Source	Cosine Similarity
tea	0.4627	fat	0.5233
vomit	0.4801	blood	0.5251
lead	0.4894	eau	0.5367
bullock	0.4930	dung	0.5421
corps	0.4934	liquid	0.5471
dust	0.5098	manure	0.5509
refuse	0.5131	**snuff**	0.5569
sick	0.5183	stomach	0.5622
sage	0.5187	beer	0.5627
bone	0.5211	**tobacco**	0.5658

Table 4: List of smell sources with cosine similarity lower than the threshold. The lower the similarity, the higher the perception shift of the smell source.

to the 1500-1799 one using a shared vocabulary. Shifts in the olfactory perception are then detected by computing the cosine similarity between the two embeddings of the same *Smell Source* in the two time periods. We focus on smell sources because we aim at analysing which items have undergone a significant change in the way their smell was perceived over time. In particular, we first compute the average similarity between all the smell sources in the two time spans and then we set as threshold for possible semantic shift the average similarity minus the standard deviation.

Table 4 shows the smell sources in the two time spans that have undergone the highest change in olfactory perception. As displayed in Table 2, the performance of the classifier is still rather law on smell sources, probably due to the evaluation strategy based on the exact match of the spans. For this reason, we performed a manual check of the smell sources detected by the shift analysis and their surrounding text. Interestingly, some items in the list were also analysed in previous historical research and were identified as key elements involved in olfactory change. We report few examples below.

Tea: The variation in the context related to this smell can be imputed to the great change in the perception of this beverage from a very exotic one when it first entered Europe (around 1630s-1640s) to a central role in the daily domestic life for the majority of Europeans (especially Dutch and English) in the nineteenth century (Webster and Parkes, 1844). Therefore, in the first period it is possible to find references to the flavor of tea as something new and not really pleasant to the European taste (see Example 1 below, extracted from our corpus), while by the early 19th Century the smell of tea becomes very common in European houses (Example 2).

(1) *Nor can it be drunk so strong without tasting an unpleasant bitterness, which the milk partly hides (1773)*

(2) *Benjamin led his mother on into the dining-room [...] the tea-table already spread, and a delicate, home-like aroma of toast and tea pervading it. (1879)*

Tobacco: Prior to the end of the 1700s, the odor of tobacco is regarded as a symbol of manliness and prevalent in most male settings. The adjectives linked to tobacco were mainly confined to the realm of male authority, with "strong" being a commonly used term to describe the scent of tobacco (Example 3). However, it isn't until the 1800s that unflattering descriptors like "disgusting", "nauseating" or "unpleasant" were linked to the scent of tobacco (Example 4).

(3) *I heard my brother say "you smell strong of tobacco". (1760)*

(4) *He had thick boorish hands, and he smelt unpleasantly of tobacco smoke. (1843)*

As smoking fell out of favor, *Snuff* emerged as the favored method of consuming tobacco (Tullett, 2019b; Goodman, 2005). Indeed, by the late eighteenth century, *Snuff* became the fashionable choice over smoking due to the prevailing manner of the period focused on the need to please others (Tullett, 2019b). Snuff's growing popularity provided in fact a more discreet form of tobacco consumption, which significantly reduced the likelihood of offending others with the pungent odor of smoke. This trend is reflected in the data through the frequent references to "pinch of snuff" and "snuff boxes" after 1800.

Even if the performance of the classifier is not very good in detecting smell sources, we consider these results promising. We plan to improve the system and to increase the amount of training data in the future to further refine our analysis.

6 Conclusions

In this paper we present the first information extraction system able to capture smell events, including smell words, smell sources and qualities. We then apply the system to four English corpora, covering a time period between 15th and 20th century.

Then, starting from the extracted data, we adapt an existing approach to semantic shift detection to capture which smell sources underwent a change in the way their odour was perceived before and after 1800. We find correspondences between the extracted items and the output of historical research concerning the smell of tobacco and tea.

Despite the limited amount of data, the results are promising and indicate that this research can yield valuable insights in the area of diachronical sensory analysis. In the future, we intend to broaden the scope of our data and conduct more comprehensive analyses on a greater variety of smell sources.

Acknowledgements

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

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Ryan Brate, Paul Groth, and Marieke van Erp. 2020. Towards olfactory information extraction from text: A case study on detecting smell experiences in novels. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 147–155, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alaa El-Ebshihy, Nagwa M El-Makky, and Khaled Nagi. 2018. Using google books ngram in detecting linguistic shifts over time. In *KDIR*, pages 330–337.

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *In Proceedings of WordNet and Other Lexical Resources Workshop*.

Roxana Girju and Charlotte Lambert. 2021. Inter-sense: An investigation of sensory blending in fiction. In *Proceedings of the 1st International Workshop on Multisensory Data Knowledge*, volume 3064. CEUR-WS.

Jordan Goodman. 2005. *"To live by smoke" in Tobacco in history: The cultures of dependence*. Routledge, pp. 212-235.

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

Stephen McGregor and Barbara McGillivray. 2018. A distributional semantic methodology for enhanced search in historical records: A case study on smell. In *Proceedings of the 14th Conference on Natural Language Processing, KONVENS 2018, Vienna, Austria, September 19-21, 2018*, pages 1–11. Österreichische Akademie der Wissenschaften.

Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroğlu, and Sara Tonelli. 2022a. Building a multilingual taxonomy of olfactory terms with timestamps. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4030–4039, Marseille, France. European Language Resources Association.

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. 2022b. 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

D. Quercia, R. Schifanella, L. Aiello, and K. McLean. 2015. Smelly maps: The digital life of urban smellscapes. In *Proceedings of ICWSM*.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice. Working paper, International Computer Science Institute, Berkeley, CA.

Strik Strik Lievers and Bodo Winter. 2018. Sensory language across lexical categories. *Lingua*, 204:45–61.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014a. A computational approach to generate a sensorial lexicon. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 114–125, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014b. Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1511–1521, Doha, Qatar. Association for Computational Linguistics.

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. 2019a. Re-odorization, disease, and emotion in mid-nineteenth-century england. *The Historical Journal*, 62(3):765–788.

William Tullett. 2019b. *Smell in Eighteenth-Century England: A Social Sense*. Oxford University Press.

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.

Thomas Webster and Mrs William Parkes. 1844. *An Encyclopaedia of Domestic Economy: Comprising Such Subjects as are Most Immediately Connected with Housekeeping...* Longman, Brown, Green, and Longmans.

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

Bodo Winter, Marcus Perlman, and Asifa Majid. 2018. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179:213–220.