A semi-supervised approach to extracting smell experiences in literature

SUBMITTED IN PARTIAL FULLFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

Ryan Brate
12888508

MASTER INFORMATION STUDIES
Data Science

FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

Date of defence: 17/07/2020

1st Examiner:
Prof. Paul Groth
Faculty of Natural Sciences,
Mathematics and Computer Science,
Informatics Institute,
University of Amsterdam

2nd Examiner:
Dr Marieke van Erp
Digital Humanities Lab,
KNAW Humanities Cluster

A semi-supervised approach to extracting smell experiences in literature

Ryan Brate
r.brate@gmail.com
University of Amsterdam

ABSTRACT

Environmental factors determine the smells we perceive, but societal factors factors shape the importance, sentiment and biases we give to them. Smell experiences in text offer a window into these factors, but they must first be identified. There is no known research into methods for extracting textual smell experiences. To this end, two variations on a semi-supervised approach to textual entity extraction have been considered in the context of smell experience identification in English literature. Textual patterns, in terms of syntax and key phrases have been assembled. The combined set of patterns from both implementations, are shown to offer significantly better precision-recall performance at the highest precision values as compared to a simple keyword search approach.

KEYWORDS

smell, identification, lexico-syntactic patterns, semi-supervised, iterative bootstrapping, English literature

1 INTRODUCTION

We rely on our senses: touch, taste, hearing, sight and smell; to complement one another in shaping our interpretation of our environment. However, if smell is a piece of the jig-saw puzzle in how we see the world, it is a piece open to individual interpretation. What we discern, and the importance we attribute to smell, is influenced by a multitude of macro and micro factors. The smells one individual finds vivid or deeply personal, conjuring memories by association, another may find innocuous.

Vroon (1997) [15] writes about the shifting historical relevance placed on smell. Its worthiness for attention, its association with social standing, lifestyle, emotion, science and superstitions, and other topical associations shifting with time. Smell experiences in text represent a bank of useful information ready to be harvested. Societal, cultural, environmental and linguistic factors shape the way that we describe the smells of our environment. Thus, conversely, our smell descriptions offer valuable insight into these factors in return. However, before such smell experiences in text can be explored, they must first be identified.

In this thesis project, we investigate *to what extent can smell experiences in English literature texts be identified using semi-supervised methods?* Specifically, focussing on smell experiences encapsulated in single sentences. To our knowledge this is the first time semi-supervised methods have been applied in identifying smell experiences in text.

The language of smell

English language vocabulary specific to the description of smell experiences is not expansive. Table 1 shows the results from a Cambridge Dictionary online [5] topic search, of words categorised as relating to smells and smelling.

Table 1: Results from Cambridge Dictionary 'smells and smelling' SMART Thesaurus search

smell-only (in all contexts)	**smell-only (in sensory contexts)**	**smell and taste only**
odour(N), odorous(A)	*fragrance(N)*	pungent(A)
malodorous(A)	**musk(N)**	pungency(N)
fetid(A), ***foetid(A)***	***fusty(A)***,*frowsty(A)*	pungently(ADV)
whiffy(A)	*ripe(A)*, *ripeness(N)*	*savour(N,V)*
smell(N,V), *scent(N)*	*reek(N,V)*, *stink(N,V)*	***acrid(A)***
smelly(A)	*stench(N)*, *niff(N)*	
scented(A)	sniff(V), **piney(A)**	
perfume(N)	waft(N,V), *stinky(A)*	
aroma(N), *aromatic(N)*	whiff(N),	
fragranced(A)		
petrichor(N)		
musty(A), **musky(A)**		

Note 1: A,N,V denotes adjectives, nouns and verbs, respectively
Note 2: underlined: words with smell strength connotations
Note 3: *italicised*: words with sentiment associations
Note 4: **bold face**: describes characteristics beyond strength or sentiment

A glance at Table 1, shows that not only is smell vocabulary in the English language limited, it is predominantly concerned with *strength* or *sentiment*. Other characteristics of smell are instead often described in terms of reference smell sources as similes.

"There is a strange unwholesome smell upon the room, **like mildewed corduroys**"

It is interesting then that despite using reference smell sources to describe the characteristics of an unseen smell, there is evidence that people are typically poor at correctly identifying common smells. In a study of a range of 80 common smells, Cain [8] reported correct identification less than 50% of the time on average. In asking the participants to re-apply the names they previously used to the same smells, the participants did so with only a 60% match rate against their original attempt.

It is reasonable to also suppose that being able to draw from a near-limitless, and sometimes subtly difference smell sources, is a source of inconsistency. Our own personal familiarity with this pool of smell sources, being a yet further source of inconsistency between individuals. Whilst a quirk of English and certainly other closely related languages, the reliance on reference smell sources is not universal. Majid et al., (2018) [14], contrasted Dutch and Jahai language speakers in describing a range of odorants. In the case of the Jahai speakers, the majority of terms used were abstract and unrelated to a source. Interestingly, the Jahai were both more consistent and greatly more controlled in the terms they used than the Dutch speaker relying on reference smell sources.

Whether the greater consistency of the Jahai is a consequence of linguistics or other factors, there is evidence that consistency in the use of smell sources in English can be conditioned. Croijmans and Majid (2016) [9] examined the accuracy and consistency of wine experts, coffee experts and people with no expertise in identifying smells. No group was better at naming smells outside of the domain of their expertise. However, it was apparent the domain experts had developed a toolkit of common smell sources they frequently drew upon, and in the case of the wine experts, they were more consistent in the application of that toolkit.

Lastly, it would also seem that people are not good at recalling smells. Brower (1947) asked a total of 152 students to imagine a variety of (common) sensory experiences. In the case of smell, only 57% could recall the smell of onions.

In short, it is clear that there is a great deal of personal subjectivity and inconsistency involved in how people describe and interpret smells and smell sources. Considering the following quote:

“The smell emanating from the lemon groves by the sea in the autumn”

The statement is grammatically simple and clearly refers to a smell. However, the descriptive element is personally subjective. Are we to interpret *‘the lemon groves’* as the smell? Is anything added to the smell characterisation by the *‘by the sea’* or the *‘in the autumn’* portion? The interpretation is dependant on the readers’ ability to recall smells from their own experience, how strongly they associate smell with the reference objects and situations.

The textual *smell experiences* to be targeted by semi-supervised methods are any instance of smell in text. However, that apparent personal subjectivity and inconsistency involved in interpreting smell must be considered when targeting such smell experiences. Consequently, this thesis project will also examine *to what extent is there agreement between people in identifying textual smell experiences?*

Semi-supervised techniques

The semi-supervised techniques to be investigated, are variants on the iterative bootstrapping technique [12]. As a semi-supervised approach, iterative bootstrapping utilises a large set of unlabelled, unstructured text, together with some manual input to direct the process. Thus, avoiding the need for large, expensive and time-consuming manually tagged datasets.

As depicted in Figure 1, iterative bootstrapping is a cyclical process. At the heart of the process is a harvesting dataset of unstructured text. The process targets some set of coincident features present in text. For example, Brin (1998) [7] employed iterative bootstrapping to target book author-title pairs. Hearst (2000) [10] employed the technique to target hypernym-hyponym pairs (e.g, car-vehicle). The target coincident features are stored in a lexicon. Extracts are assembled from the harvesting dataset, where they match a lexicon entry. Patterns in terms of syntax and key phrases, which target the appropriate features, are then manually determined from the extracts. The lexicon is updated with new entries, based on pattern matches and retrieval of the appropriate features, according to the determined patterns. The process can be seeded by introducing known features into the lexicon, or known patterns into the pattern set, as show in Figure 1. Both Brin and Hearst, used this process to assemble a lexicon, the pattern set being effectively a by-product of the process. However, for our purposes, it is the set of patterns that are of interest, and their potential use to identify smell experiences.

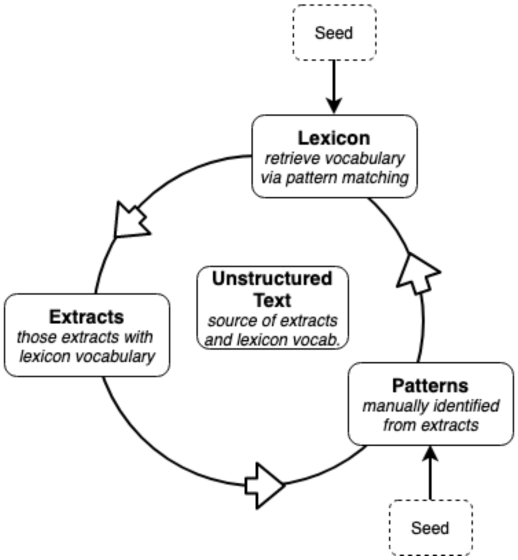
Seed
Lexicon
retrieve vocabulary via pattern matching
Unstructured Text
source of extracts and lexicon vocab.
Extracts
those extracts with lexicon vocabulary
Patterns
manually identified from extracts
Seed

Figure 1: Basic Iterative Bootstrapping process

The textual features targeted by Brin and Hearst represented distinct real world concepts, or entities, linked through a conceptual relationships. The notion of smell does not conform to a natural set of paired entities reflecting a relationship in the same way as authors and book titles do. However, there are several parts of speech which help characterise the smell experience and can therefore be targeted as features. We have already discussed the importance of smell sources. Noun groups representing smell sources being referenced are then obvious candidates for targeting. Similarly, other parts of speech, such as verbs and adjectives have been selected as features for targeting. These features are highlighted in the example, below. Note that in the example, ‘fragrance’ and ‘smell’ despite being nouns related to smell, are not reference smell sources and are therefore not targeted.

“An $\underline{\text{odd}}_{\text{adj}}$ fragrance, a smell of $\underline{\text{damp plaster}}_{\text{ref. smell source}}$, $\underline{\text{wafted}}_{\text{verb}}$ from the new house to his senses”

There remains the question of how these features should be targeted as coincident combinations. I.e., pairs, a triplet or individually, since, unlike the aforementioned lexicon pairs of Hearst and Brin, there is no inherent relationship defining the number of coincident features necessary. Instead it is a question of how restrictive we wish to make the criteria for matching sentences in the harvesting dataset. The greater number of coincident features, the fewer extracts we can expect to retrieve. Thus, a too restrictive choice may result in stalling of the iterative bootstrapping process. Conversely, if the choice is not restrictive enough and too many extracts unrelated to smell are returned, the process becomes uninformative. Iatropoulos et al., (2018) [11] in evaluating smell related vocabulary, concluded that the most commonplace words used in smell descriptions, are those which could apply to a wide range of sensory contexts. This makes intuitive sense given the heavy reliance on reference smell sources to define smell characteristics. Hence, single features are not targeted, as being to relevant outside of smell contexts. In this pilot study, pairs of coincident features have been considered for iterative bootstrapping.

The feasibility of each stage of each iterative bootstrapping cycle is critical to the success of the overall approach. Thus, in support of the main research question of this thesis project, the following supplementary research questions will also be addressed directly:

To what extent can lexico-syntactic patterns be employed to identify smell extracts and target lexicon features?

To what extent can new extracts be bootstrapped from lexicon entries?

2 RESEARCH DESIGN

Variations on iterative bootstrapping implemented

Two implementation variations on iterative bootstrapping have been investigated, each targeting two distinct sets of coincident features. A basic assumption is being made that the coincidence of the targeted features from a smell extract, is indicative of another extract being a smell experience when these features are present. The targeted features of the implementations are as follows:

(1) **Adjectives** modifying the smell experience perception, and the coincident **noun group** acting as the reference smell (one or more nouns modified by adjectives);

    E.g., *'An **odd** fragrance, a smell of **damp plaster**, wafted from the new house to his senses'*,where 'odd' and 'damp plaster' are the adjective and noun group, respectively.

    Thus, it is assumed that 'odd' and 'damp plaster', being coincident in defining this smell experience, are indicative of a smell experience when both present in other extracts.

(2) A **noun group** acting as the reference smell, and the coincident **verb group** (adverbs, verbs, associated prepositions) describing how the smell moves;

    E.g., 'An odd fragrance, a smell of **damp plaster**, **wafted from** the new house to his senses', where 'damp plaster' and 'wafted from' are the noun group and verb group, respectively.

    Thus, it is assumed that 'damp plaster' and 'wafted from', being coincident in defining this smell experience, are indicative of a smell experience when both present in other extracts.

In addition to supporting the research questions previously posed, considering two variations on the same methodology allows us to consider:

Does targeting different feature pairs result in the identification of pattern sets which target different smell experiences?

Datasets

A collection of English literature texts was assembled from Project Gutenberg [3]. Iterative bootstrapping benefits from a higher quantity of available smell extracts. New patterns are discoverable, only when new extracts are identified and retrieved, based on features identified from seen extracts. The greater the number of relevant extracts, the greater the opportunity for pattern variety. Also, a large number of extracts offers a certain redundancy, meaning patterns missed in one cycle may be picked up in another. Accordingly, the texts have been selected on the basis of the highest rate of occurrence of keywords derived from Table 1. To facilitate an automated pattern matching approach, each sentence in this collection has been independently (syntactically) parsed via the spaCy [4] toolkit. Each word in a sentence has been tagged with its dependency tag and parts of speech (POS) tag. For example, the parsed form of the text, below, in Listing 1:

"The tempting aroma of the precious wine seemed to mingle with the soft strange words"

```
"det_the_DET amod_tempting_ADJ conj_aroma_NOUN
↪  prep_of_ADP det_the_DET amod_precious_ADJ
↪  pobj_wine_NOUN conj_seemed_VERB aux_to_PART
↪  xcomp_mingle_VERB prep_with_ADP det_the_DET
↪  amod_soft_ADJ amod_strange_ADJ pobj_words_NOUN"
```

Note: Prior to parsing, each sentence undergoes a text analysis stage of tokenising via NLTK [1] and lower-casing
Note: Refer to https://spacy.io/api/annotation for the meaning of the tags

Listing 1: Example of a parsed dataset entry

This text collection from Project Gutenberg was split into 3 separate datasets:

- A harvesting dataset of 99 texts;
- A validation dataset of 20 texts;
- An evaluation dataset of 20 texts.

Creating the gold-standard

From the evaluation dataset a gold-standard of manually labelled sentences has been assembled:

- Seven documents, each of 100 (randomly assigned) sentences and annotated once by a single annotator. There are four annotators in total, including the author;
- One additional document, of 100 (randomly assigned) sentences, annotated independently by three annotators. None of the annotators for this set is the author.

Despite the texts of the harvesting, validation and extract sets being chosen for their high frequency of Table 1 derived keywords, on average only approximately 1 in 100 sentences contain a keyword. Thus, assuming that smell experiences typically contain a keyword, the evaluation set contains smell experiences in very low proportion. A gold standard set of extracts has been sampled from the evaluation dataset such to ensure a substantial number of smell extracts. The evaluation set has been scanned for the high smell association keywords derived from Table 1. 80% of the sentences in the gold standard documents contain a Table 1 related word, the remaining 20% having been randomly sampled. There is no overlap between documents, or redundancy within a document.

Annotators have highlighted and annotated spans according to the following criteria [2]:

- 'd'. A smell description;
  E.g., 'An odd fragrance, a smell of damp plaster, wafted from the new house to his senses'
  The inherent subjectivity of when precisely a smell experience becomes a description is left to the perception of the annotator.
- 'o'. A smell alluded to without expansion of its characteristics;
  E.g., 'A fragrance wafted from the new house to his senses'.
- 'v'. Any verb in the sentence which is associated with smell generally or with a specific smell experience within the extract;
  E.g., 'An odd fragrance wafted from the new house to his senses'
- 's'. Sense of smell alluded to directly;
  E.g., 'An odd fragrance, wafted from the new house to his senses'

Additionally, two documents (of the aforementioned group of 7) have been annotated with an additional set of tags:

- 'a'. An adjective being applied to the smell alluded to;
  E.g., 'An odd fragrance, a smell of damp plaster, wafted from the new house to his senses'
- 'n'. The noun group referred to as a smell source; E.g., 'An odd fragrance, a smell of damp plaster, wafted from the new house to his senses'

Answering the research questions

To support answering the main research question of this thesis project, several supporting questions are posed. These supporting questions are listed as follows, together with the approaches taken to resolve them.

Table 2: Number of text spans by annotation tag

Annotation tag	Number of corresponding text spans
'd'	432
o'	112
'v'	147
's'	28
'a'	37
'n'	75

(1) *To what extent is there agreement between people in identifying textual smell experiences?*

The inter-annotator agreement is considered with regards the single gold standard document, which has been subject to independent annotation by multiple annotators. Specifically:

- The level of agreement with respect to those sentences which have been tagged as *a smell experience*, by one or more annotators. I.e., those sentences with a text spans annotated with either 'd' or 'o';
- The level of agreement with respect to those sentences which have been tagged as *a smell description*, by one or more annotators. I.e., those sentences with a text span annotated with 'd';
- The level of agreement in respect of those verbs within sentences which have been tagged with 'v', i.e., verbs the annotator perceives as being related to smell generally, or in the context of the sentence.

(2) *To what extent can lexico-syntactic patterns be employed to identify smell extracts and target lexicon features?*

- With respect to the gold standard set, and where a smell experience is an extract with a span labelled 'd' or 'o'. The precision-recall characteristics of the identification patterns returned from iterative bootstrapping are compared against a simple keyword search. These keywords are words with high smell association, derived from Table 1, replicated in Appendix A. Thus, evaluating the usefulness of the returned patterns as compared to a simple more obvious approach.
- The targeted lexicon features were specific in their relation to smell, e.g., a noun group acting as the reference smell, not just a noun in a textual smell experience. This is an important distinction, since the introduction of unrelated vocabulary into the lexicon, risks reducing the quality of the corresponding harvesting extracts. A portion of the gold standard set has been annotated with 'v', 'a' and 'n', i.e., those verbs, adjectives and noun groups targeted as features by the iterative bootstrapping implementations. Using these documents, the level of agreement between the gold standard and the patterns in targeting these features is considered.

(3) *To what extent can new extracts be bootstrapped from lexicon entries?*

This is answered with respect to the record of new lexicon entries, new extracts and number of resulting patterns per cycle.

(4) *Does targeting different feature pairs result in the identification of pattern sets which target different smell experiences?*

The recall of the pattern sets of implementation 1 and implementation 1 and 2 combined, in identifying *true* smell extracts in the gold standard set, has been compared via a paired t-test.

Pattern Representation

A high level pattern representation approach was adopted, to capture, and enable pattern matching with, parts of speech, synonyms and flexible groupings of these in text. This is very important given the enormous inherent flexibility of language. Thus enabling fewer, relatively simple, user-defined patterns to match against a greater pool of textual smell experiences that do not differ substantially. The basic building block of these high-level representations is the parsed word format exampled in Listing 1, i.e., *dependency_word_POS*. E.g., *_smell_* would match against any instance of the parsed form of the text 'smell' regardless of dependency and POS tag, whereas *_smell_VERB* would only match with the verb form.

Words derived from Table 1 with high smell-association, feature frequently as *key words* or in *key phrases* in identified patterns and were assembled in synonym groups. For example, *<smell_noun>* is defined to match against the parsed forms of 'aroma', 'odour', 'scent', 'perfume' etc. That is, those words that in very many contexts may be used interchangeably with the noun 'smell'. Similarly, adjective and verb synonym groups of the keywords derived from Table 1 were assembled.

The POS representations, or chunks, define various possible arrangement of the respective POS, resulting from listing and optional supporting POS. For example, the *<verb>* chunk is able to match against both 'singing' and 'loudly and raucously singing, laughing and dancing wildly'. These chunks were defined, and updated in process, based on observed language patterns from harvesting set extracts.

```
[<adj>] <smell_noun> _,_* _of_ <pronoun>* [<noun> {_of_ <noun>}*]
```

Listing 2: Example identified pattern, expressed

Listing 2 is an example of a high-level pattern representation identified in implementation 1. It will match against the following two example extracts taken from the harvesting set, extracting the targeted adjective and noun group reference smell source(s) pair:

"the **warm** aroma of **multitudinous exotics**"

"the **ammoniacal** smell of **the horses**"

Specifics of the Bootstrapping Implementation

Figure 2 shows the intermediary stages of the bootstrapping process as implemented. The main outcomes of the process, a lexicon of targeted coincident features, and a pattern set are highlighted in blue. The differences in the implemented process of Figure 2 and the general methodology of Figure 1 are: The addition of a validation loop; and the discrimination between identification and extraction patterns, explained as follows.

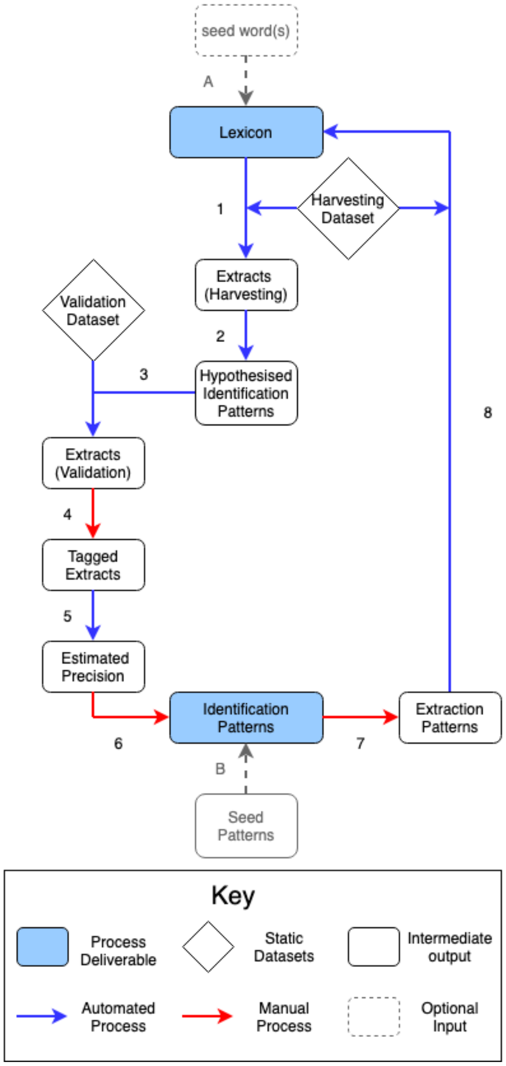

Figure 2: Implemented Iterative Bootstrapping Process

Note: refer to the implementation source code [6]

Description of the process.

- **Seed Word(s):** In both variants the lexicon has been seeded with the high smell association word *_aroma _NOUN*. With reference to Table 3, _aroma_NOUN was selected since it results in a very manageable number of corresponding extracts.

Table 3: Comparison of number of extracts matched with seeds _aroma_NOUN and _smell_NOUN

	seed word	No. corresponding extracts
1	_aroma_NOUN	91
2	_smell_NOUN	1131

- (A) **Lexicon:** A collection of the coincident features pattern matched from the harvesting set (or seed word), stripped of dependency tags and determiners;
  E.g., [‘_tempting_ADJ’, ‘_precious_ADJ _wine_NOUN’]

- (1) **Extracts (Harvesting):** Previously unseen extracts (parsed and original form) which match any of the latest cycle of lexicon entries in full;
  E.g., The‘**precious wine** exuded a **tempting** perfume

- (2) **Hypothesised Identification Patterns:** The harvested extracts are examined in-turn, and new patterns are hypothesised. For example, based on the preceding example extract, we may hypothesis a pattern:
  <noun> exuded __DET <adj>* <smell_noun>
  note: * denotes zero or more consecutive <adj> occurrences

- **Validation Loop:** Patterns with low precision risk introducing a large volume of vocabulary into the lexicon which is unrelated to smell. The validation loop is an attempt to limit this, by estimating pattern precision and setting a minimum acceptance threshold.
  - (3) **Extracts (Validation):**: A random sample, from the validation dataset, of 10 extracts per hypothesised pattern;
  - (4) **Tagged Extracts:**: The validation extracts are manually tagged as TP (True Positive), FP (False Positive) or U (unknown);
  - (5) **Estimated Precision:** An estimate of precision of each hypothesised is performed, ignoring extracts tagged as unknown;

$$\frac{TP}{TP + FP} \tag{1}$$

  Those hypothesised patterns that pass a validation threshold of 0.7 estimated precision are accepted. If no example of a pattern is present in the validation set, it is passed.

- (6,7) **Identification Pattern and Extraction Pattern Assembly:** Two variations of the hypothesised patterns are retained, identification patterns and extraction patterns.
  Extraction patterns target the previously discussed feature pairs, such to introduce new vocabulary into the lexicon. Thus, extraction patterns are used to drive each iterative cycle. For example, based on the preceding hypothesised pattern
  E.g., ‘[<noun>] exuded __DET [<adj>] <smell_noun>’

  Identification patterns are a superset of the extraction pattern set, and are concerned with matching any and all smell experiences, not just matching feature pairs. The identification pattern set, then, is our desired output from the iterative bootstrapping process.
  E.g., ‘<noun> exuded __DET <adj>* <smell_noun>’

- (8) **Repopulating the lexicon:** The extraction patterns are applied to the harvesting set, and targeted entities are collected as coincident groups and added to the lexicon.

Four complete cycles of Implementation 1 and three complete cycles of Implementation 2 were completed.

3 RESULTS

Implementation 1 outcomes

Following some rationalisation of the identification patterns resulting from the iterative bootstrapping implementation, 103 identification patterns are determined.

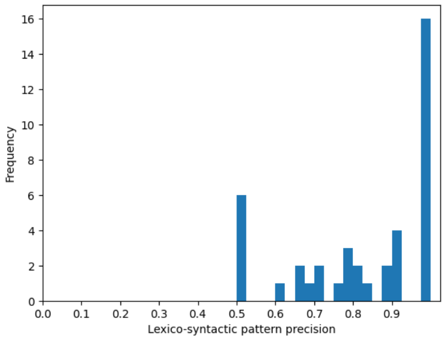

Figure 3: Histogram of Implementation 1 pattern set precisions with respect to the gold standard

The overwhelming majority of these patterns involve Table 1 derived words, collected under synonym groups as discussed in Section 2. For example:
where * denotes zero or more consecutive occurrences:

- ‘<adj>* compound__ <smell_noun>’
  E.g., ‘a delightful forest **aroma**

- ‘<adj> _with_ __DET* <pronoun>* <smell_noun> _of_ <pronoun>* <verb> <noun> {_of_ <noun>}*’
  E.g., ‘heavy with the **smell** of freshly turned soil’

There were additionally a small number of patterns identified that do not involve the Table 1 vocabulary, as follows:

- <adj>* _breath|breaths_ _of_ <pronoun>* <noun> {_of_ <noun>}*
  E.g., ‘...and inhale the sweet breath of autumn, which was borne upon gentle gales’

- <adj>* _breeze_ _washes_ _in_ <pronoun>* <noun> {_of_ <noun>*}
  E.g., ‘the soft warm breeze washes in dark fruit, dark flowers...’

- <adj>* _exhalation|exhalations_ _of_ <noun> {_of_ <noun>}*
  E.g., 'leaving a faint exhalation of scent and powder and delicate perfumes'

- <adj>* _air_ _was|is_ _laden_ _with_ {__DET* <smell_noun> _of_}+ <pronoun>* <noun> {_of_ <noun>}*
  E.g., 'The air was laden with orange and chocolate

- '_air_* _,_ _sweet_ _with_ <pronoun>* <noun> {_of_ <noun>}*'
  E.g., ' the mild air, sweet with fading leaves and bracken'

Implementation 2 outcomes

After a rationalisation process, a total of 55 identification patterns are determined. 54 of the 55 identified patterns, involve Table 1 derived words, collected under synonym groups as discussed in Section 2. E.g.,

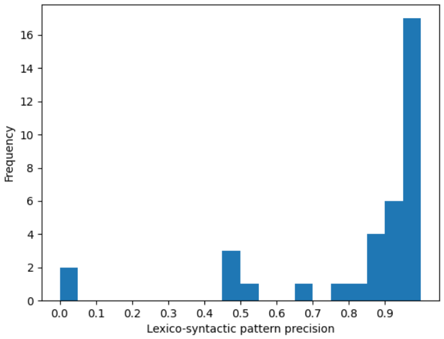

Figure 4: Histogram of Implementation 2 pattern set precisions with respect to the gold standard

- '<smell_noun> _of|like_ __DET* <pronoun>* **<noun> {_of_ <noun>}* <verb> prep__**'
  E.g., 'the aroma of **new-sawn timber and sawdust mingled with**...'

The single example of lexico-syntactic pattern not involving Table 1 derived words, uses a noun (incense) as synonymous with smell.

- '_fumes_ _of_ _incense_ {_of_ <noun>}* __DET* <verb> prep__*'
  E.g., 'the heavy fumes of incense rose up'

Addressing the research questions

Results are presented in terms of the supporting research questions, posed previously.

To what extent can new extracts be bootstrapped from lexicon entries? Table 4 and Table 5, show the statistics of each cycle for implementations 1 and 2, respectively. The columns in bold show the number of extracts returned in each cycle. Evidently given the successive cycles of each implementation, extracts conforming to previously unseen patterns were successfully bootstrapped.

As previously acknowledged in Section 1, smell relies on a vocabulary which can similarly be applied in other sensory contexts. To streamline each cycle's extract returns, only unseen lexicon pair entries, unseen extracts, and extracts not conforming to a previously identified pattern were retrieved at each cycle. Despite this, it is apparent from Table 4 and Table 5, that the number of returned extracts spiralled beyond that which allows for manual inspection. Several cycles relied on Appendix A key word search to target extracts for examination. It was the author's experience, that the extracts often exhibited very low densities of smell experiences.

Table 4: Record of iterative cycles outcomes for Implementation 1: targeting coincident adjectives modifying the smell, and noun group reference smells

Cycle	Lexicon entries	New (unseen) extracts	Hypothesised (new) patterns	New id. patterns/ New ex. patterns
0	1**	**91**	15	15 / 13
1	519	**1,509**	28	26 / 22
2***	874	**4,216**	14	13 / 8
3	463	**464**	4 ****	4 / 4

**Seed word: _aroma_NOUN

*** sifted with Table 1 word search due to high volume

**** not subject to validation as cycles stopped

Note 1: Each lexicon (pair) entry is unique, and each extract is unique

Table 5: Record of iterative cycles outcomes for Implementation 2: targeting coincident verb groups associated with the smell experience and noun group reference smells

Cycle	Lexicon entries	New (unseen) extracts	Hypothesised (new) patterns	New id. patterns/ New ex. patterns
0	1**	**91**	11	10 / 9
1***	530	**2,968**	12	10 / 9
2	565	**1,030**	11 ****	11 / 8

**Seed word: _aroma_NOUN

*** sifted with Table 1 word search due to high volume

**** not subject to validation as cycles stopped

Note: Each lexicon (pair) entry is unique, and each extract is unique

Does targeting different feature pairs result in the identification of pattern sets which target different smell experiences? Figure 5 shows the relative precision-recall performance of the group predictions, in respect of the gold standard, with regards the pattern sets of: implementation 1; implementation 2; and implementation 1 and 2 combined.

It is apparent that the implementation 1 pattern set (red) consistently outperformed the pattern set of implementation 2 (black); and that the combined pattern sets of implementations 1 and 2 (green), generally outperform the pattern set of implementation 1 alone. With reference to Appendix B, Table 8, a significance test has been performed comparing the relative performance of the combined pattern set, with that of implementation 1. The significance test confirms with a 5% significance level, where patterns have a precision greater than 0.7, that the combined pattern set *significantly* has superior recall performance at corresponding precision cut-offs. Thus, we can conclude that targeting different feature pairs did

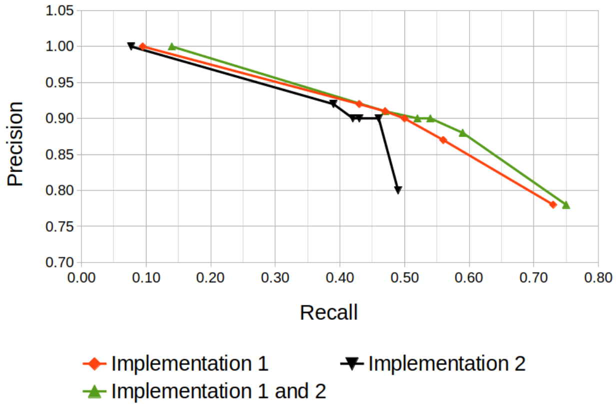

Figure 5: Group prediction precision-recall performance of the pattern sets of: implementation 1; implementation 2; and implementation 1 and 2 combined.

result in patterns which target *significantly* different smell extracts in the gold standard set.

To what extent is there agreement between people in identifying textual smell experiences? Cohen's Kappa is a metric used to measure pairwise inter-annotator agreement. A Cohen's Kappa of 0, denotes an even probability of agreement. Landis and Koch (1977) [13] denote a Cohen's Kappa score of 0.41 to 0.60, and .61 - 0.80 as representing *moderate* and *substantial* strength of agreement, respectively. Classifying a score of 0.81 to 1.0 is as near perfect agreement.

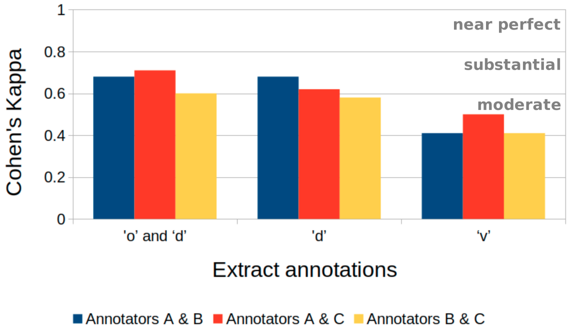

Figure 6: Cohen's Kappa scores of pairwise annotator agreement

Figure 6 shows the pairwise annotator agreement with regards the single gold standard document of 100 extracts, annotated by multiple annotators. All annotators are in *substantial* agreement in identifying all and any extracts that allude to smell, i.e., all spans labelled 'o' or 'd'. Annotators are generally in substantial agreement in identifying extracts which *describe* smell experiences, i.e., extracts with a spans labelled 'd'. Although, one pair of annotators are at the very upper end of *moderate* agreement only. Finally, in identifying verbs either highly associated with smell, or associated with smell in the context, there was only *moderate* annotator agreement.

It is reasonable to assume that human error, i.e., misreading, miscomprehending or simple skipping an extract, played some role in the observed imperfect inter-annotator agreement scores. Instances of likely human error are apparent on inspection of the gold standard, in those instances where there is arguably little room for personal subjectivity. E.g., as follows, for which one of the three annotators did not attribute either a 'd' or 'o' tag to the span.

"There was a smell of decaying leaves and of dog."

However, a number of extracts clearly demonstrate the potential for subjectivity in smell experience interpretation, as a source for annotator disagreement:

E.g., In the following extract, each of the three annotators attributed 'd', 'o' and no tag to it, respectively.
"Seated beside her aromatic rest, In silence musing on her loveliness, Her knight and troubadour."

E.g., In the following extract, one of three annotators tagged it as 'o', the other two presumably thought it sufficiently descriptive to be tagged 'd'.
"Between each pair of columns an elegant table of cedar bore on its platform a bronze cup filled with scented oil, from which the cotton wicks drew an odoriferous light."

To what extent can lexico-syntactic patterns be employed to identify smell extracts and target lexicon features? Figure 7 shows the precision-recall group performance with regards the gold standard, comparing a simple keyword search (blue) to the combined implementation 1 and 2 pattern sets (green). The keyword search terms are as per Appendix A. The graph was produced by considering a series of lower bound precision cut-offs. Only those patterns whose precision exceeded the cut-off were included in the group. With a group identifying an extract (rightly or wrongly) when one or more patterns (or keyword) in the group predicts matches with it.

It is important to note that it was beyond the scope of this thesis project to produce an *exhaustive* set of patterns, with each implementation being halted at a few cycles. Thus, the precision recall-curve performance of the patterns sets from implementation 1 and 2, of Figure 7, represent a *lower-bound*, with respect to the potential performance of iterative bootstrapping. Further identified patterns, from additional iterative bootstrapping cycles, would only increase recall or do nothing, but not decrease recall at precision cut-offs. Hence region A, where the keyword search shows superior precision-recall performance, over the lower-bound performance of the combined patterns of implementations 1 and 2, is less informative.

Region B, however, shows a trend at the very highest precision levels, of superior precision of the combined pattern sets of implementation 1 and 2, as compared to keyword search, for the same recall. With reference to Appendix B, a one-tailed paired t-test

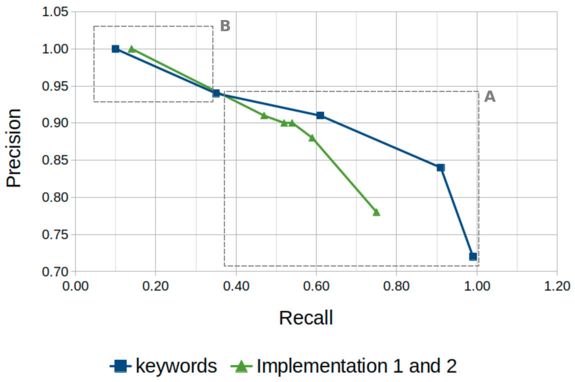

Figure 7: Group prediction precision-recall performance of the pattern sets of: *Implementation 1*; *Implementation 2*; *Implementation 1 and 2 combined* and a *Table 1 derived keyword search*

shows the differences to be statistically significant to a 5% significance level.

Addressing the second aspect of the research question: the performance of the resulting patterns in targeting lexicon features. Table 6 and Table 7, show the confusion matrices with respect to the portion of the gold standard where those adjective, noun and verb groups targeted in the implementations have been annotated. Based on the confusion matrices, Cohen's Kappa scores of 0.64 and 0.43 have been calculated for implementation 1 and 2, respectively. That is, implementation 1 is in *substantial* agreement with annotators in correctly identifying targeted features, whereas implementation 2 is only in *moderate* agreement.

	pattern prediction		
		TRUE	FALSE
gold standard	TRUE	9	5
	FALSE	4	184

Table 6: Confusion matrix with respect to the Implementation 1 pattern set in extracting *true* feature pairs

	pattern prediction		
		TRUE	FALSE
gold standard	TRUE	10	7
	FALSE	14	172

Table 7: Confusion matrix with respect to the Implementation 2 pattern set in extracting *true* feature pairs

4 DISCUSSION AND CONCLUSION

This thesis project is concerned with the question, *to what extent can smell experiences in English literature texts be identified using semi-supervised methods?* The implementation of two variations on iterative bootstrapping have demonstrated that semi-supervised techniques can be used to determine textual patterns, which can be used to identify smell experiences in text. Whilst the overwhelming majority of identified patterns involved keywords and phrases with a high smell association, the implementations revealed a number of new phrases used in smell contexts. Furthermore, it has been shown that at the very highest levels of precision, pattern group identification of smell experiences offers significantly better recall rates than a keyword search.

The focus of application of iterative bootstrapping implementations was centred on single sentence extracts specifically in regards to English literature texts. It would be interesting to explore the applicability of this semi-supervised method to other textual contexts, and longer-distance relationships spanning multiple sentences. Additionally, it would be informative to explore the influence of tweaking the implementation's parameters and approaches, with regards the precision-recall performance of the resulting pattern sets, such as:

- Explore the influence of different seed words on the process outcomes, and consider sequential re-seeding strategies;
- Investigate the impact of a higher validation precision threshold with regards the number, and quality, of extracts returned each cycle and the corresponding identified patterns;
- Investigate the impact of changes to the pattern chunk definitions in terms of smell experience return precision and recall.
  E.g., 'ripe' and 'smelly' have a 9.5% and 99% keyword search precision rate on identifying extracts, yet the _smell_ADJ synonym group which includes both, makes no distinction between them.
  E.g., the pattern segment *<pronoun>* <noun> {_of_ <noun>}**, is used to match noun groups used as reference smells. However, it was defined on an ad-hoc basis only, based on extract observations.
- Complete the iterative cycles of each implementation to exhaustion, enabling a more complete precision-recall evaluation;
- Explore the impact of targeting a noun group, verb group, adjective feature triplet, rather than feature pairs.

The primary benefit of a semi-supervised approach to identifying language patterns, rather than trawling through text line by line, is the promise of significantly minimised user effort. The implementations demonstrated the potential utility of the resulting patterns in identifying textual smell experiences. However, the number of extracts, and the quality of extracts in terms of smell experience density was identified as source of inefficiency which would benefit from being addressed further. The implementation statistics of Table 4, for example, show the explosion in the number extracts for manual examination, in certain cycles, which correspond to only comparatively few new patterns being identified.

The observed high volume of low smell experience extracts may be an inherent challenge of smell experiences relying on vocabulary which is equally, or more so, applicable in other sensory contexts. I.e., the targeting and adding to the lexicon of words with a low smell association, resulting in poorer quality extracts. However, there are a number of clear, possible avenues to explore to improve the smell association of words added to the lexicon:

- Increasing of the validation set size, ensuring more accurate precision estimates. Also, disregarding patterns for which we have no validation set examples, and using a range of higher precision cut-offs for inclusion as *extraction patterns*. Thereby improving the level of smell association of lexicon entries on average;
- Explore the effects of using more coincident features simultaneously, i.e., pairs were selected on the basis that if one feature alone was weakly associated with smell, two together may improve the association. More coincident features may further improve the likelihood of an extract relating to smell.

Finally, consideration as to degree of agreement between people in their interpretation, suggested a less than perfect agreement not only of the subtly nuanced aspects of smell experiences, but even at recognition of smell experiences as a broad classification. On inspection of annotations, however, it is unclear how many of these were genuine discrepancies in terms of subjective perceptions. More annotators, supported by a more comprehensive approach to tagging, e.g., requiring the annotators to explicitly note their reasoning and deliberations would help. This would offer a window into the mind of the annotators, reinforcing any conclusions that may be drawn.

In summary, the main conclusions of this thesis project are:

- Semi-supervised methods can be used to successfully identify textual smell extracts in English literature. It has been shown that the targeting of noun groups, verb groups and adjectives used in smell, can be used to successfully bootstrap new, unseen extracts. Furthermore, it has been demonstrated that the resulting patterns can result in *significantly* greater recall at higher precision than a keyword search approach. However, extension of this research would be required to more fully determine the precision-recall benefits of identified patterns, over cruder approaches.
- A significant challenge of applying iterative bootstrapping to textual smell experience identification, is the applicability of the targeted vocabulary to other sensory contexts. This has been shown to have the potential to substantially increase the volume and reduce the density of smell experiences in harvesting extracts.
- The assessment of agreement between people in regards to their perception of smell experiences showed imperfect agreement. This may be the case, and is supported by inherent subjectivity involved in interpreting textual smell experiences. However, further data is needed for a better understanding, of where instances of annotator disagreement are due to subjectivity and where they are due to human error.

Whilst there is scope for further study, the experiments in this thesis project have shown that smell experiences can be identified, and smell experience features can be extracted from text. The source code and supporting materials are available on github [6].

A TABLE 1 DERIVED KEYWORDS

acrid, aroma, aromas, aromatic, bouquet, fetid, foetid, fragrance, fragrances, fragranced, frowsty, fusty, malodorous, musk, musky, musty, niff, niffs, odorous, odour, odours, olfaction, perfume, perfumes, perfumed, petrichor, pong, pongs, piny, piney, pungency, pungent, pungently, putrid, redolence, redolent, reek, reeks, reeked, ripe, ripeness, savour, scent, scents, scented, smell, smells, smelled, smelt, smelly, sniff, sniffs, sniffed, stench, stink, stinks, stinky, waft, whiff, whiffs, whiffy

B SIGNIFICANCE TESTS

General Methodology

All significance tests performed in this thesis project are concerned with significant differences in recall performance at difference precision cut-offs. Recall performance is only concerned with *true* smell experiences in the gold standard. Hence, in considering whether there is a significant difference in recall, the significance tests are comparing pattern set matches against the true smell experience set.

Group predictions correctly identifying (i.e., matching against) an extract as a smell experience are allocated the number 1: failure to identify, are allocated the number 0. Thus the paired differences in prediction performance is given by $group_1 - group_2$:

E.g., paired differences = [1, 0, 0, 1, -1, ...]

Null hypothesis assumption: mean(paired differences) = 0
i.e., each group is equally performant in identifying smell experience extracts. The hypothesis tests are conducted on the basis of a 5% significance level.

The difference in the group identification performance of the pattern sets of Implementation 1 and Implementation 1 and 2 combined

Table 8 shows the p-value results of the significance test for a one-tailed paired-test, based on an alternative hypothesis of mean(paired differences) > 0. I.e., that the pattern set of implementation 1 and 2 combined, is better at correctly identifying true smell experiences than the pattern set of implementation 1 alone.

Table 8: P-values for one-tailed paired t-test

precision cut-off	mean difference	st. dev	p value
0.99	0.049	0.22	1.1E-6
0.95	0.049	0.22	1.1E-6
0.90	0.040	0.20	9.4E-6
0.85	0.045	0.21	1.8E-6
0.80	0.040	0.20	9.4E-6
0.70	0.020	0.25	0.047
0 - 0.60	0.011	0.23	0.15

Note: data size of 452 smell experience extracts

Thus with respect to Table 8, for precision cut-offs considered greater or equal to 0.7, the null hypothesis is rejected and the alternative hypothesis accepted.

The difference in the group identification performance of the pattern sets of implementation 1 and 2 combined and Appendix A keywords

Table 9 shows the p-value results of the significance test for a one-tailed paired-test, based on alternative hypothesis of mean(paired differences) > 0. I.e., that the pattern set of implementation 1 and 2 combined, is better at correctly identifying true smell experiences than a keyword search.

Table 9: P-values for one-tailed paired t-test

precision cut-off	mean difference	st. dev	p value
0.99	0.042	0.46	0.027
0.95	0.042	0.46	0.027

Note: data size of 452 smell experience extracts

Thus, at precisions of greater than or equal to 0.95, the null hypothesis is rejected and the alternative hypothesis accepted.

REFERENCES

[1] https://www.nltk.org/. accessed: 2020-02-13.

[2] Annotation guidelines. https://github.com/ryanbrate/DS_thesis/tree/master/6_Evaluation/annotation%20guidelines. Accessed: 2020-07-10.

[3] Free ebooks - project gutenberg. https://www.gutenberg.org/. Accessed: 2020-03-05.

[4] Industrial strength natural language processing in python. https://spacy.io/. accessed: 2020-03-05.

[5] Smells & smelling. https://dictionary.cambridge.org/topics/senses-and-sounds/smells-and-smelling/. Accessed: 2020-03-05.

[6] Thesis project source code. https://github.com/ryanbrate/DS_thesis. Accessed: 2020-07-10.

[7] Sergey Brin. Extracting patterns and relations from the world wide web. 06 1998.

[8] W S Cain. To know with the nose: keys to odor identification. *Science (New York, N.Y.)*, 203(4379):467–470, 1979.

[9] Ilja Croijmans, Asifa Majid, and Sidney Arthur Simon. Not all flavor expertise is equal: The language of wine and coffee experts. *PLoS ONE*, 11(6):e0155845, 2016.

[10] Marti Hearst. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conference on Computational Linguistics (CoLing)*, 05 2000.

[11] Georgios Iatropoulos, Pawel Herman, Anders Lansner, Jussi Karlgren, Maria Larsson, and Jonas K. Olofsson. The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. *Cognition*, 178:37 – 49, 2018.

[12] Rosie Jones, Andrew Mccallum, Kamal Nigam, and Ellen Riloff. Bootstrapping for text learning tasks. 08 1999.

[13] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[14] Asifa Majid, Niclas Burenhult, Marcus Stensmyr, Josje de Valk, and Bill S. Hansson. Olfactory language and abstraction across cultures. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, 373(1752):20170139–20170139, 2018.

[15] Amerongen A. . Vroon, P. A. and H. Vries. *Smell: The secret seducer.* Farrar, Straus and Giroux, 1997.