A semi-supervised approach to extracting smell experiences in literature

Ryan Brate

University of Amsterdam

r.brate@gmail.com

17-07-2020

Overview

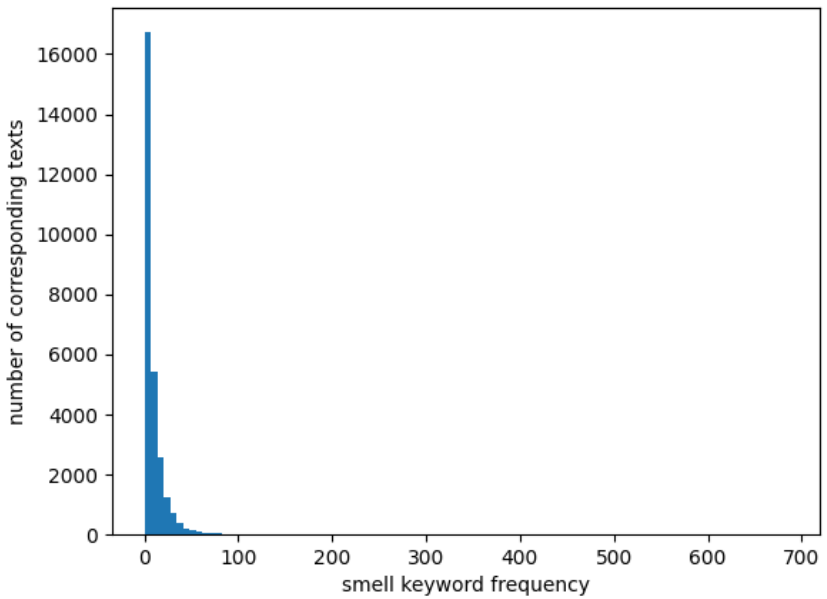
Introduction

"To what extent can smell experiences in English literature texts be identified using semi-supervised methods?"

Data

unstructured text

Source: Project Gutenberg [1], ≈ 30,000 English language texts



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 from

Assembled Collection: 139 texts

Divided in 3 sets: **Harvesting**, **Validation** and **Evaluation** set

Gold Standard

assembly

- 8 documents, each of 100 extracts
  - 80% based on high smell association keywords, 20% random sample

Gold Standard

assembly

- 8 documents, each of 100 extracts
  - 80% based on high smell association keywords,
    20% random sample
    - 'd'
      E.g., 'An odd fragrance, a smell of damp plaster, wafted from the new house to his senses'
    - 'o'
      E.g., 'A fragrance wafted from the new house to his senses'.

Gold Standard

assembly

- 8 documents, each of 100 extracts
  - 80% based on high smell association keywords, 20% random sample
    - 'd'
      E.g., 'An odd fragrance, a smell of damp plaster, wafted from the new house to his senses'
    - 'o'
      E.g., 'A fragrance wafted from the new house to his senses'.
    - 'v'
      E.g., 'An odd fragrance wafted from the new house to his senses'
  - 2/8 documents additional tags
    - 'a', 'n'
      E.g., 'An odd$_a$ fragrance, a smell of damp plaster$_n$, wafted from the new house to his senses'

Gold standard

supporting questions

1: "To what extent is there agreement between people in identifying textual smell experiences?"

Gold standard

supporting questions

1: "To what extent is there agreement between people in identifying textual smell experiences?"

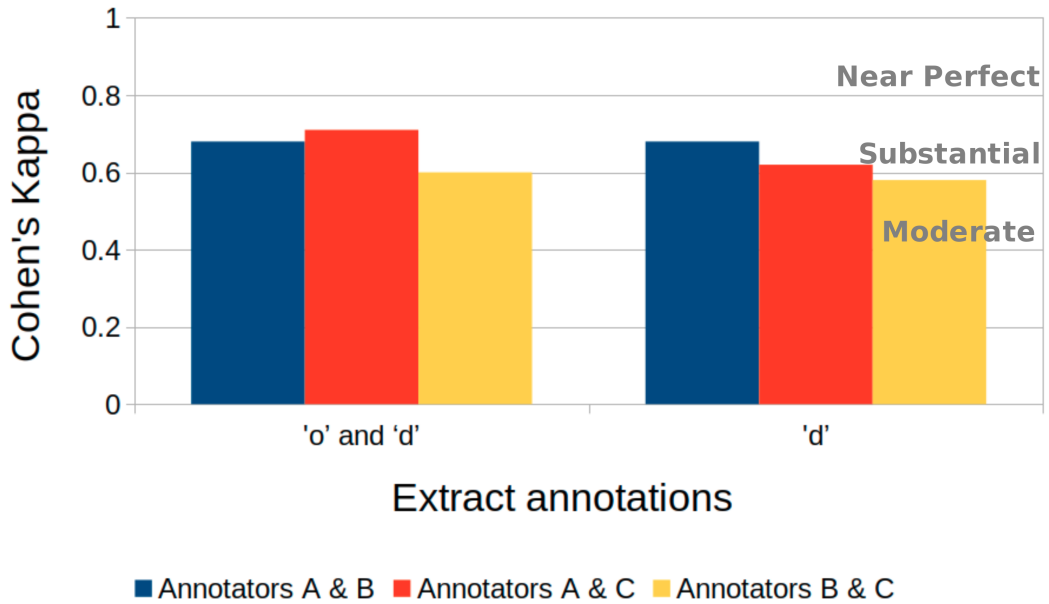


Figure: Cohen's Kappa scores of pairwise annotator agreement

Gold Standard

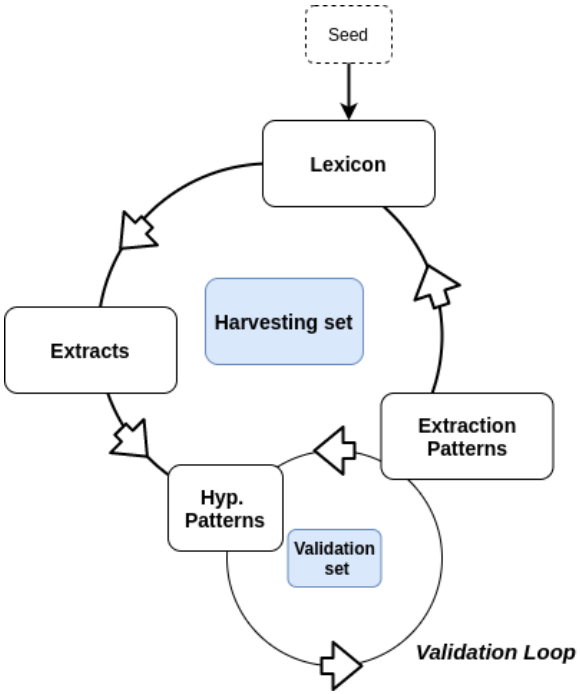
Supporting Questions

E.g., disagreement ‘o’ vs. ‘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.”

E.g., Disagreement as to whether an allusion to smell at all
And do you know, my charming young lady, and you, my generous protector, do you know, you who breathe forth virtue and goodness, and who perfume that church where my daughter sees you every day when she says her prayers?—For I have brought up my children religiously, sir.”,

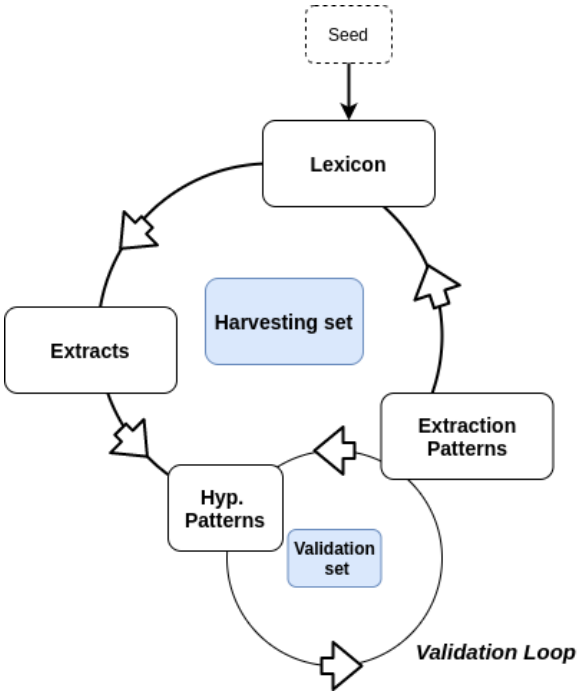
Iterative Bootstrapping

Example based on Hearst (2000) [3], targeting hypernym-hyponym pairs



Iterative Bootstrapping

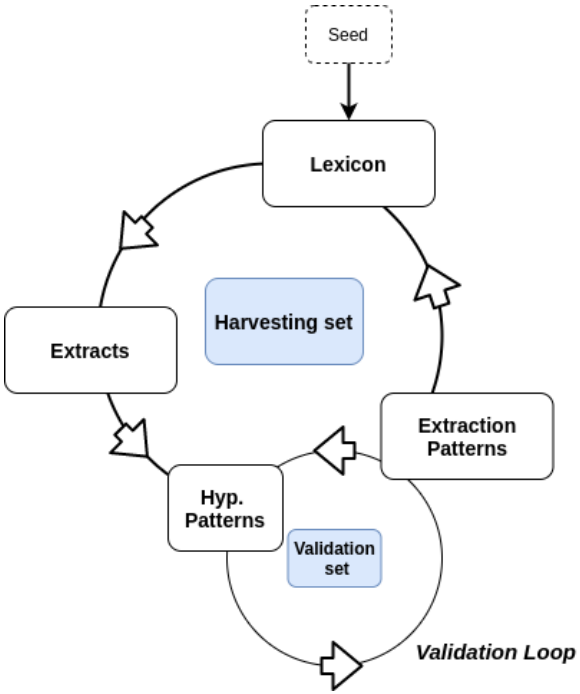
Example based on Hearst (2000) [3], targeting hypernym-hyponym pairs



- **Lexicon:**
  E.g.,
  '*European Countries*', '*The Netherlands*'

Iterative Bootstrapping

Example based on Hearst (2000) [3], targeting hypernym-hyponym pairs

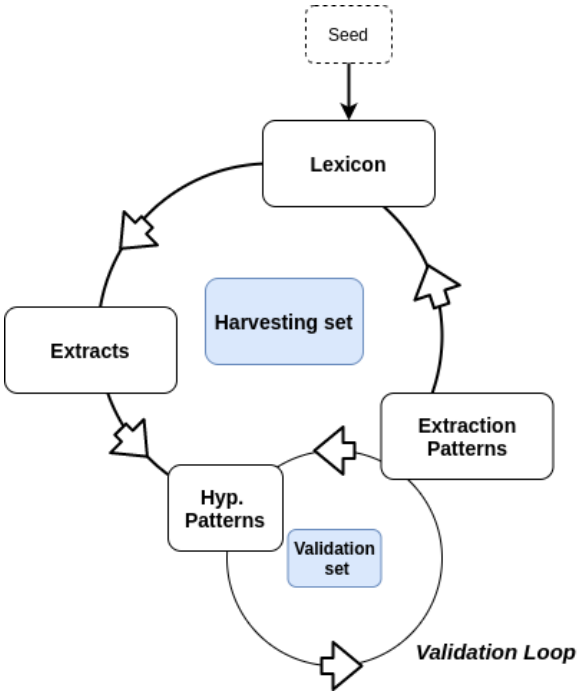


- **Lexicon:**
  E.g.,
  '*European Countries*', '*The Netherlands*'

- **Extracts:**
  'Amongst *European Countries*, *the Netherlands* has the greateast ratio of bikes per person'

Iterative Bootstrapping

Example based on Hearst (2000) [3], targeting hypernym-hyponym pairs



- **Lexicon:**
  E.g.,
  '*European Countries*', '*The Netherlands*'
- **Extracts:**
  'Amongst *European Countries*, *the Netherlands* has the greateast ratio of bikes per person'
- ****Patterns**:**
  'amongst *Noun*, *Noun*'

Iterative Bootstrapping Adaptation

targeted features

Gold standard extract example:

'd' : 'An odd fragrance, a smell of damp plaster, wafted from the new house to his senses'

Iterative Bootstrapping Adaptation

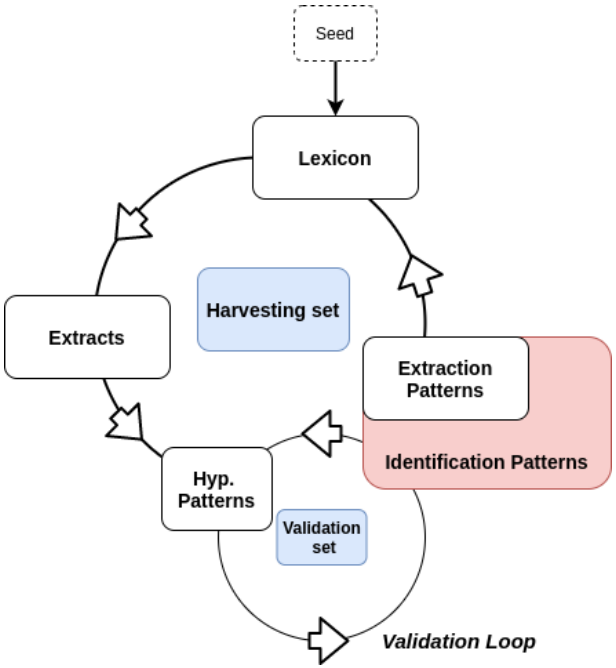
targeted features

Gold standard extract example:

'd' : 'An odd$_{adjective}$ fragrance, a smell of damp plaster$_{noun\ group}$, wafted$_{verb\ group}$ from the new house to his senses'

Iterative Bootstrapping Adaptation

implemented process



Implementation 1:

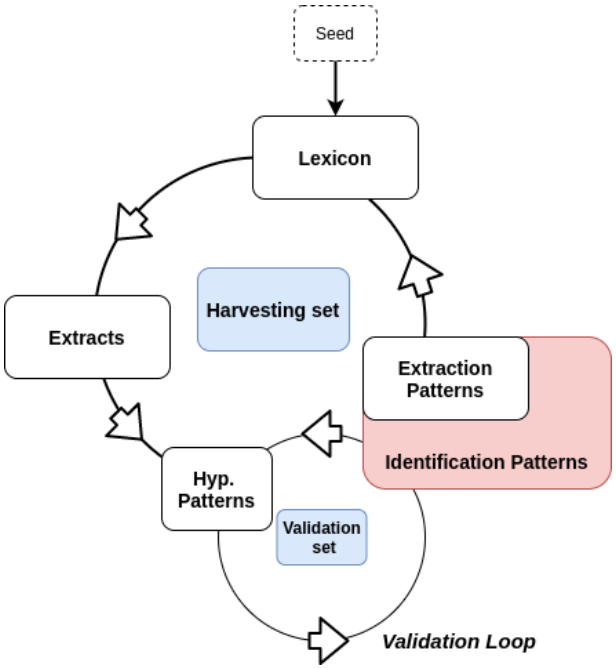
- target NOUN groups + ADJ

Implementation 2:

- target NOUN groups + VERB

Iterative Bootstrapping Adaptation

implemented process



Implementation 1:

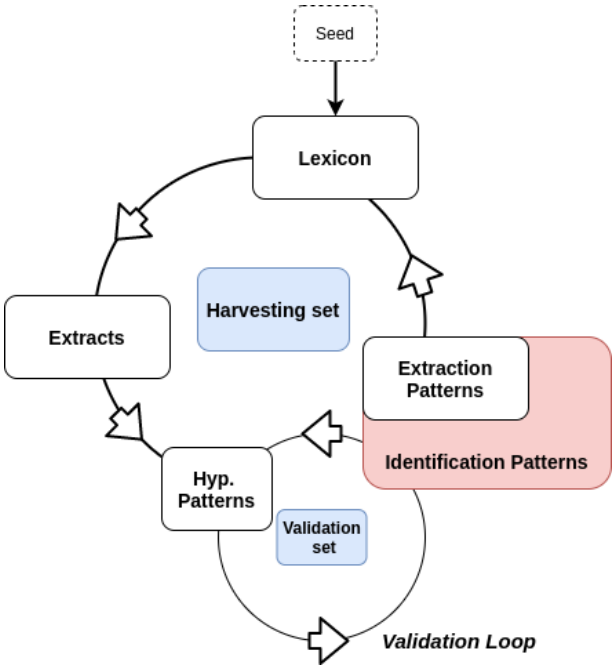
- target NOUN groups + ADJ
- seed with _aroma_NOUN

Implementation 2:

- target NOUN groups + VERB
- seed with _aroma_NOUN

Iterative Bootstrapping Adaptation

implemented process



Implementation 1:

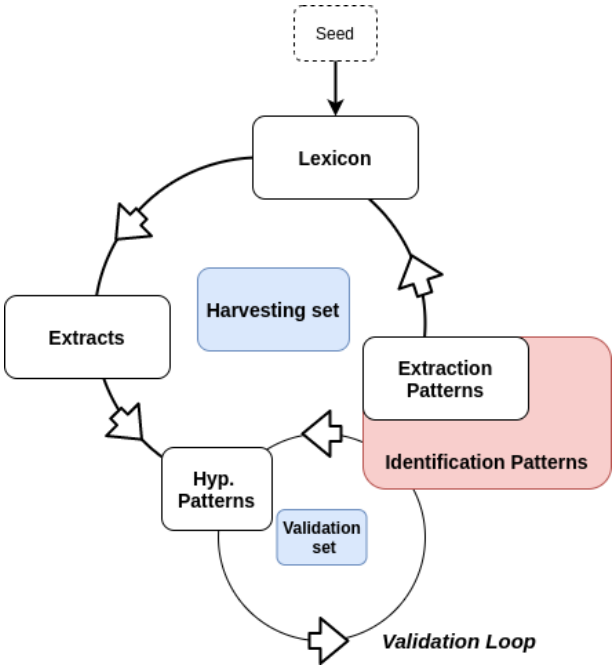
- target NOUN groups + ADJ
- seed with _aroma_NOUN
- validation threshold of 0.70

Implementation 2:

- target NOUN groups + VERB
- seed with _aroma_NOUN
- validation threshold of 0.70

Iterative Bootstrapping Adaptation

implemented process



Implementation 1: *:

- target NOUN groups + ADJ
- seed with _aroma_NOUN
- validation threshold of 0.70
- 4 cycles completed

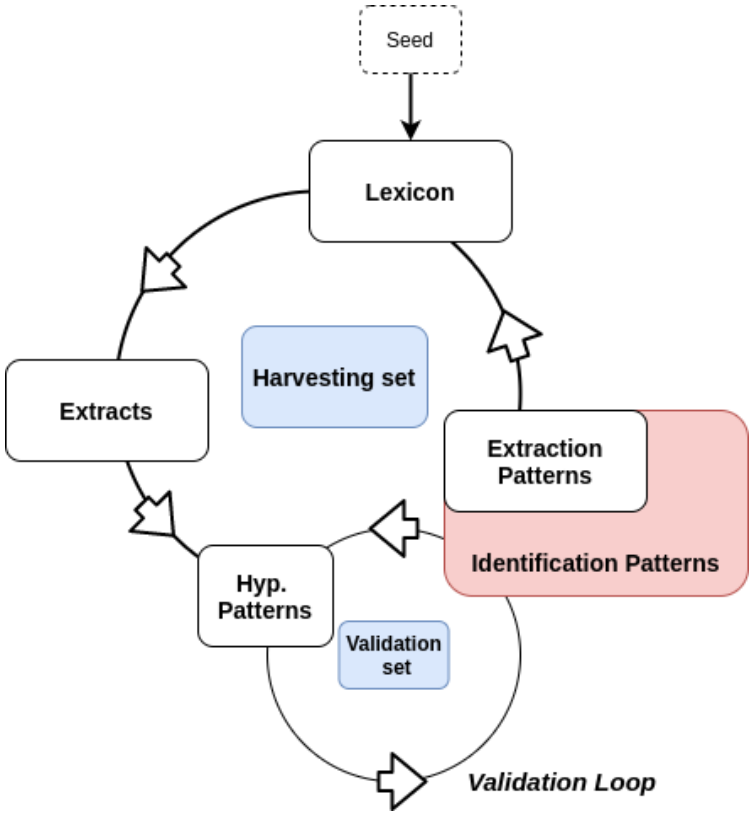
Implementation 2:

- target NOUN groups + VERB
- seed with _aroma_NOUN
- validation threshold of 0.70
- 3* cycles completed

halted at 3 cycles, since, since stat., significant results observed

Iterative Bootstrapping Adaptation

implemented processes



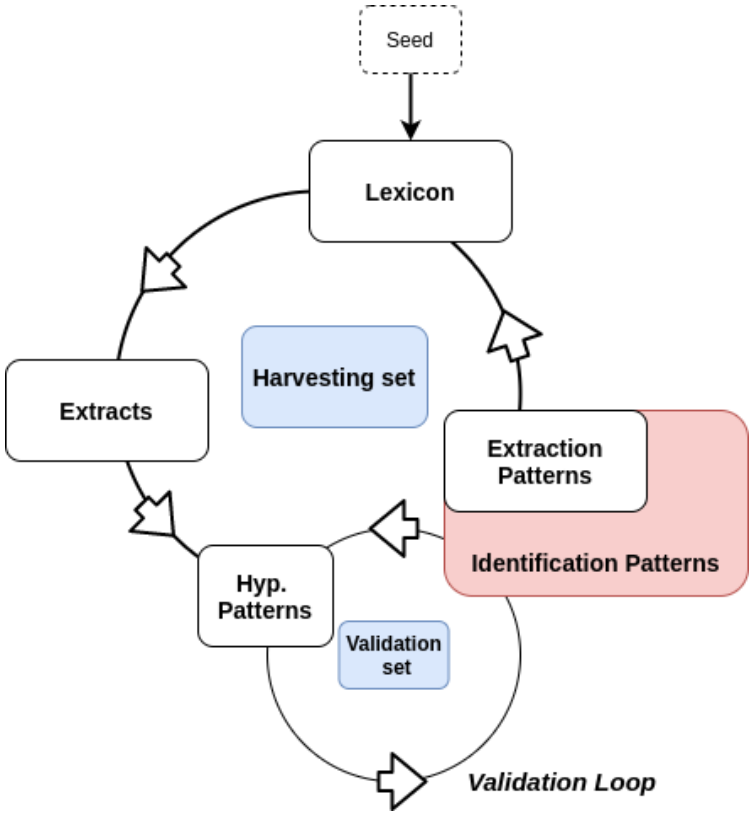
Extraction Patterns:

The sweet$_{adj}$ violets$_{noun}$ lent fragrance

ID Patterns:

Iterative Bootstrapping Adaptation

implemented processes



Extraction Patterns:

The sweet_{adj} violets_{noun} lent fragrance

ID Patterns:

The sweet_{adj} violets_{noun} lent fragrance

The sweet_{adj} fragrance

The fragrance of violets_{noun}

Iterative Bootstrapping Adaptation

pattern matching

How do we target, Adjectives, Nouns, Verbs?

Pattern Representation

Iterative Bootstrapping Adaptation

pattern matching

How do we target, Adjectives, Nouns, Verbs?

spaCy [2] : *dependency_text_POS*

Pattern Representation

Iterative Bootstrapping Adaptation

pattern matching

How do we target, Adjectives, Nouns, Verbs?

spaCy [2] : *dependency_text_POS*

E.g.,
The fragrance of violets$_{noun}$

det_The_DET ROOT_fragrance_NOUN prep_of_ADP pobj_violets_NOUN

Pattern Representation

Iterative Bootstrapping Adaptation

pattern matching

How do we target, Adjectives, Nouns, Verbs?

spaCy [2] : *dependency_text_POS*

E.g.,
The fragrance of violets$_{noun}$

det_The_DET ROOT_fragrance_NOUN prep_of_ADP pobj_violets_NOUN

Pattern Representation

$< adj >$: "$\{\_and|, |or\_^*\_\_ADJ + \_and|, |or\_^*\}^+$
Distinctive, strong and unpleasant

Iterative Bootstrapping Adaptation

pattern matching

How do we target, Adjectives, Nouns, Verbs?

spaCy [2] : *dependency_text_POS*

E.g.,
The fragrance of violets$_{noun}$

det_The_DET ROOT_fragrance_NOUN prep_of_ADP pobj_violets_NOUN

Pattern Representation

$< adj >$: "$\{\_and|, |or\_^*\_\_ADJ + \_and|, |or\_^*\}^+$
Distinctive, strong and unpleasant

$< smell\_noun > \_as\_ < adj > \_of|like\_ < noun >$

Results

Supporting question:

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

Results

Supporting question:

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

Identified Patterns:

$< adj > * < smell\_noun >$
$\_of|like\_ < pronoun > * <$
$verb >< noun >$
$\_of\_ < noun > *\}$
...

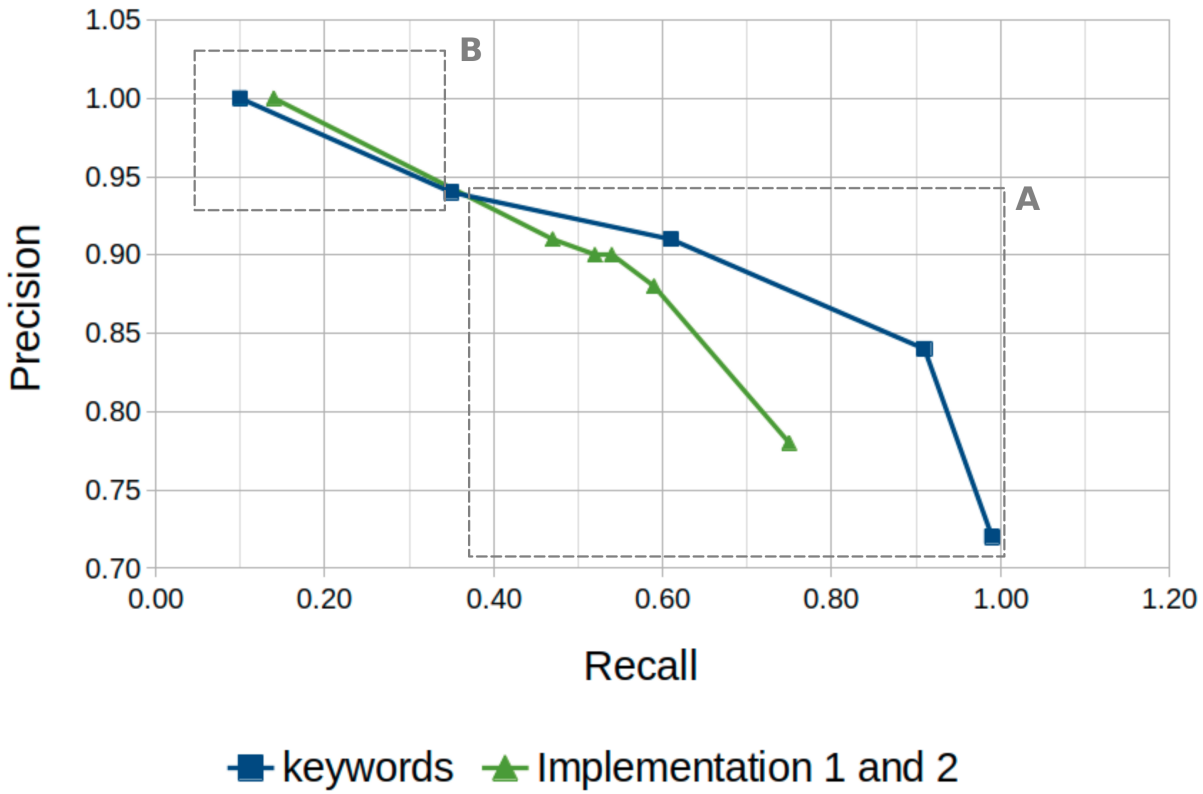
vs

Reference Case:

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

Results

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



Results

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

	pattern prediction		
gold standard		TRUE	FALSE
	TRUE	9	5 (FN)
	FALSE	4 (FP)	184

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

	pattern prediction		
gold standard		TRUE	FALSE
	TRUE	10	7 (FN)
	FALSE	14 (FP)	172

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

Results

Examples of FN in implementation 2:

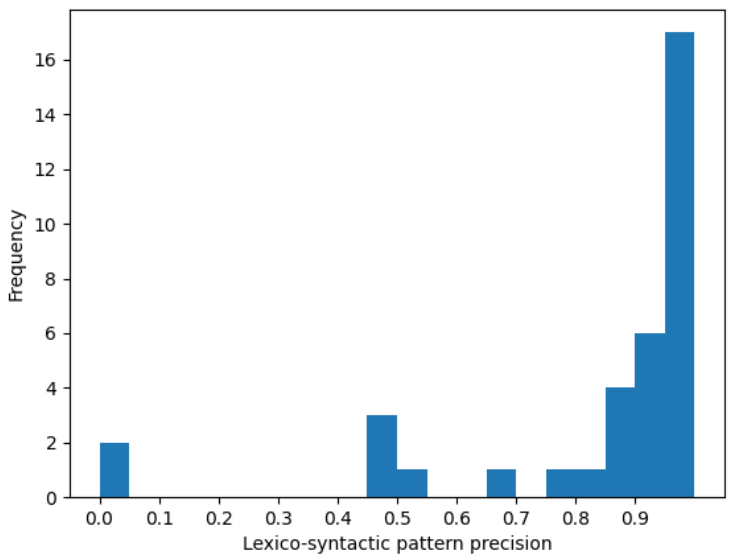
1. Incorrect parsing:
   "A faint perfume stole to his nostrils"

   det_A_DET amod_faint_ADJ nsubj_perfume_NOUN **ROOT_stole_NOUN** prep_to_ADP poss_his_DET pobj_nostrils_NOUN punct

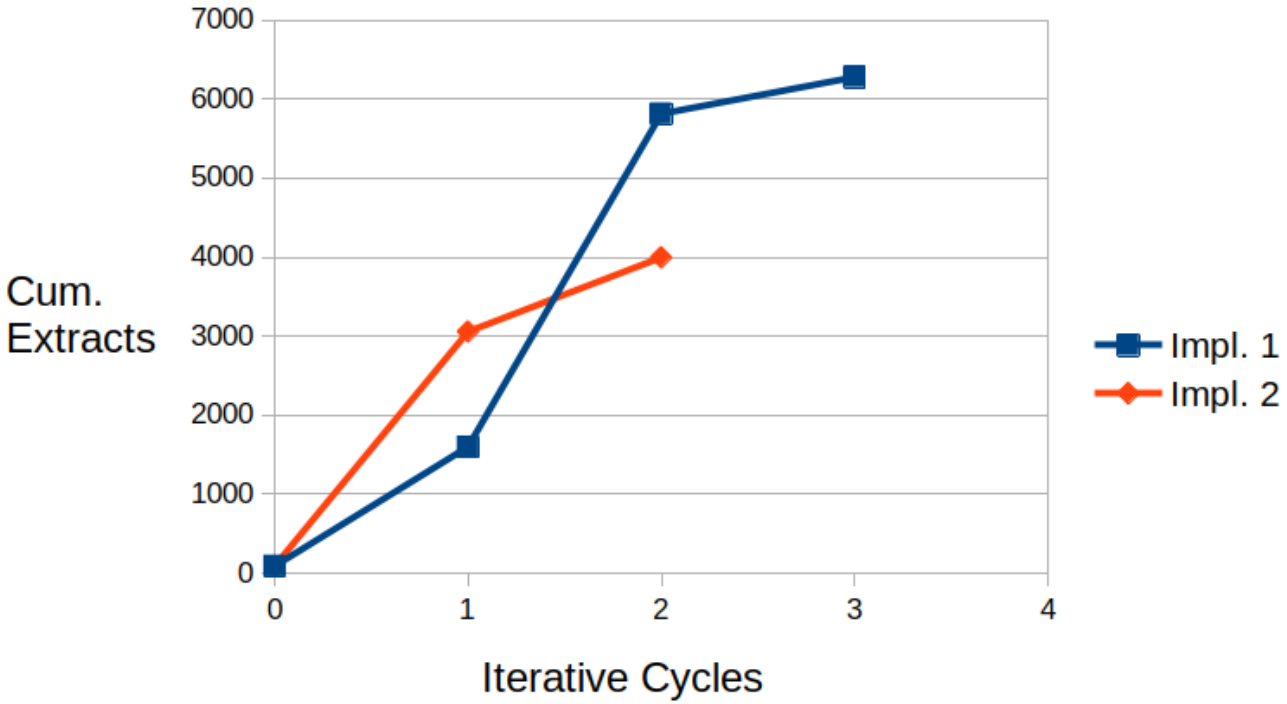
2. No matching pattern:
   "the fresh scent of the grass"
   $< adj >^* < smell\_noun > \_of\_ < noun >$ (present in impl. 1)

Results

- Examples of FP in implementation 2:
  "the 7th of june tattered"
  $[< noun > \{\_of\_ < noun >\}*]\_, \_[< verb > prep\_\_*] < noun >$



Results


7000
6000
5000
4000
3000
2000
1000
0
Cum. Extracts
0
1
2
3
4
Iterative Cycles
Impl. 1
Impl. 2

Results

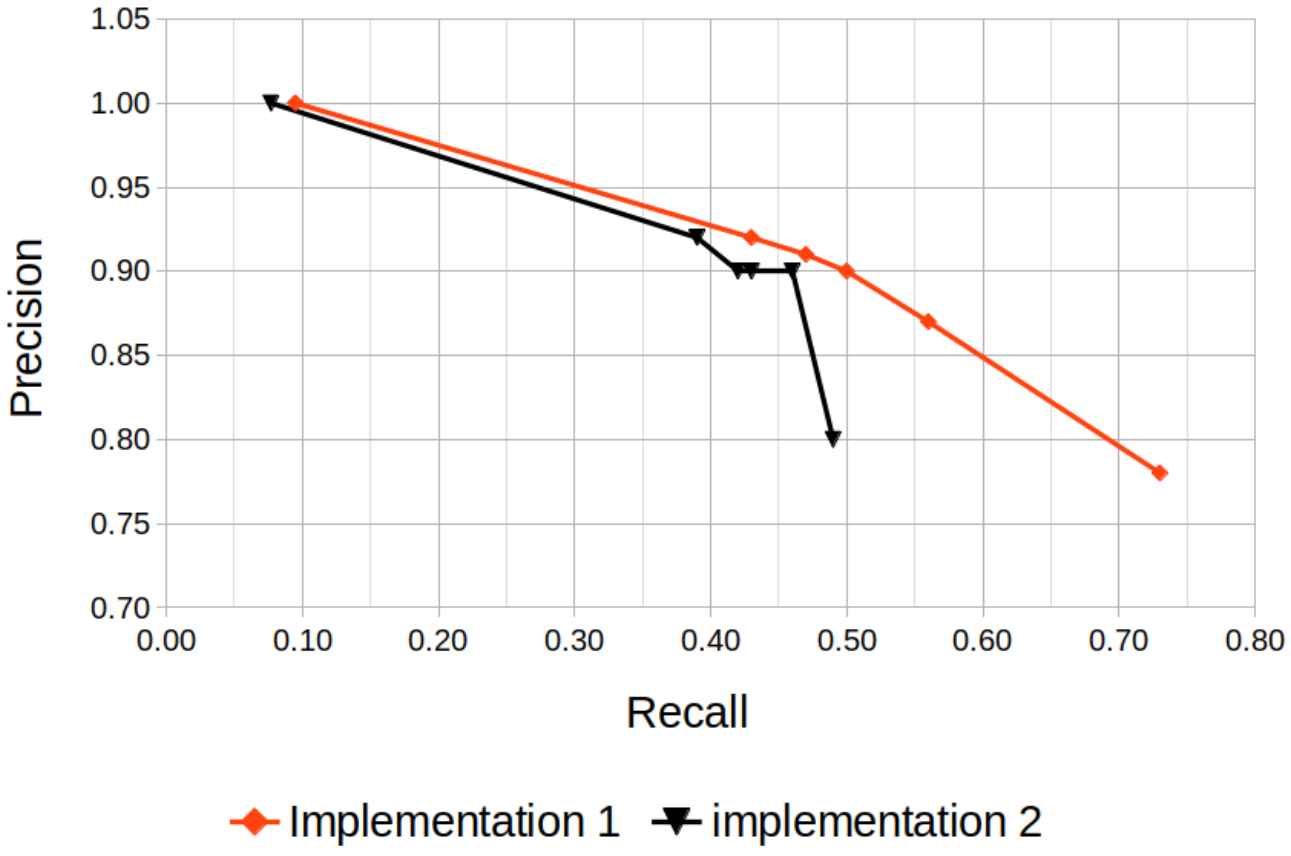
Supporting question:

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

Results

Supporting question:

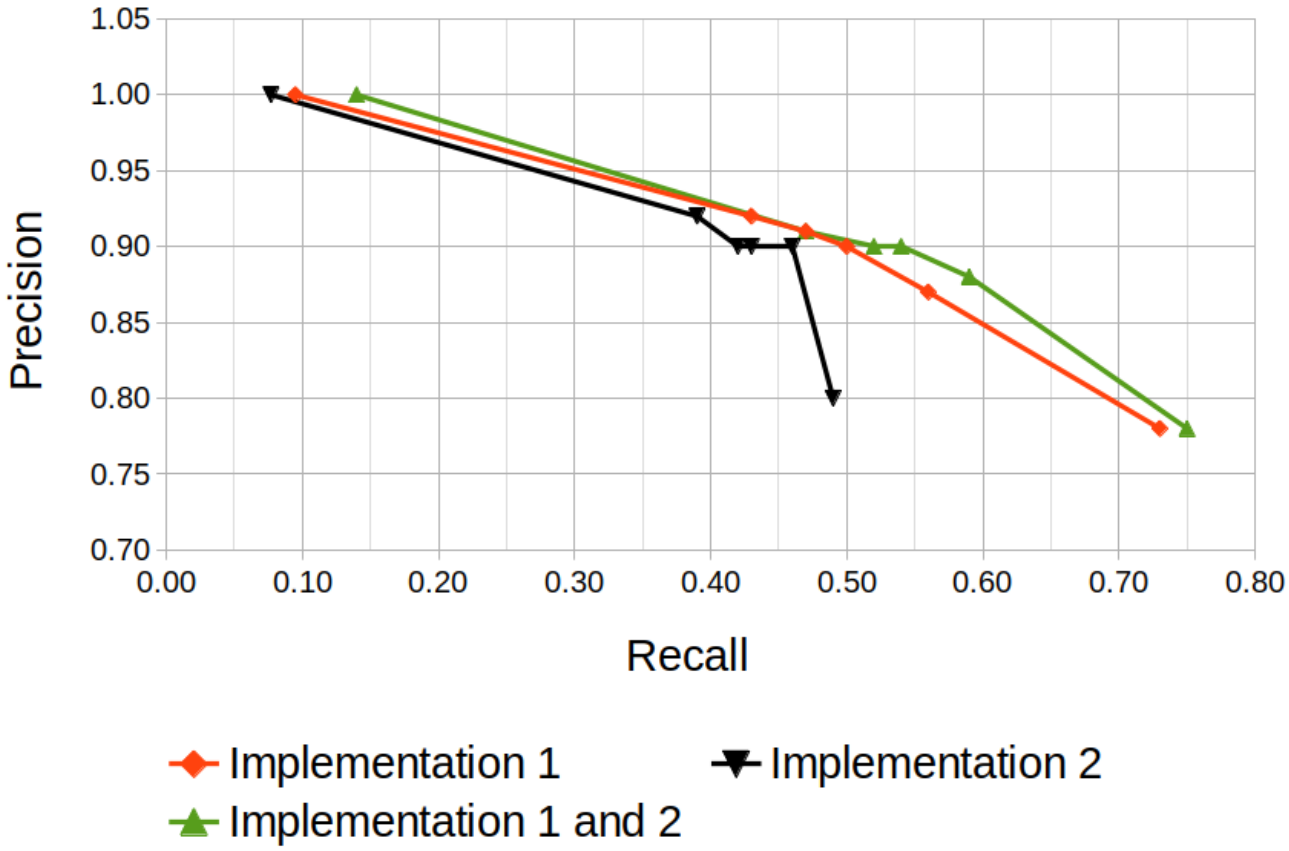
3: "Does targeting different feature pairs result in the identification of pattern sets which target different smell experiences?"



Results

Supporting question:

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



Conclusion

To what extent can smell experiences in English literature texts be identified using semi-supervised methods?

- Semi-supervised methods targeting language features can be used to identify smell experiences.
  (The experiment was successful)
- The resulting patterns demonstrated superior recall at higher precision values than keyword search.
- Revealed extract examples that would not be picked up with a keyword search.
- There is further work to be done to refine the implementation

Further Work

Parameters:

- Validation threshold, more fine-grained synonym groups
- Consider the impact of seed words, seeding strategies
- Explore variations on pre-defined language chunks, synonym groups

General Approach

- Target verbs, adj, noun simultaneously
- Complete the each iterative bootstrapping implementation to exhaustion

The End

References

Free ebooks - project gutenberg.

`https://www.gutenberg.org/`.

Accessed: 2020-03-05.

Industrial strength natural language processing in python.

`https://spacy.io/`.

accessed: 2020-03-05.

Marti Hearst.

Automatic acquisition of hyponyms from large text corpora.

Proceedings of the 14th Conference on Computational Linguistics (CoLing), 05 2000.

title

To what extent can smell experiences in English literature texts be identified using semi-supervised methods?

To what extent is there agreement between people in identifying textual smell experiences?

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?

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