Integrating Query Performance Prediction in Term Scoring for Diachronic Thesaurus

Chaya Liebeskind and Ido Dagan

Bar-Ilan University

LaTeCH 2015

Research Context:
Domain Specific Diachronic Corpus

Example: searching **vegetarian** in biblical scholarship archive

modern texts with references to ancient language

Were All Men Vegetarians before the Flood?
…God instructed Adam saying, "I have given you every herb that yields…" (Genesis 1:29) …

(by Eric Lyons, M.Min.)

ancient texts

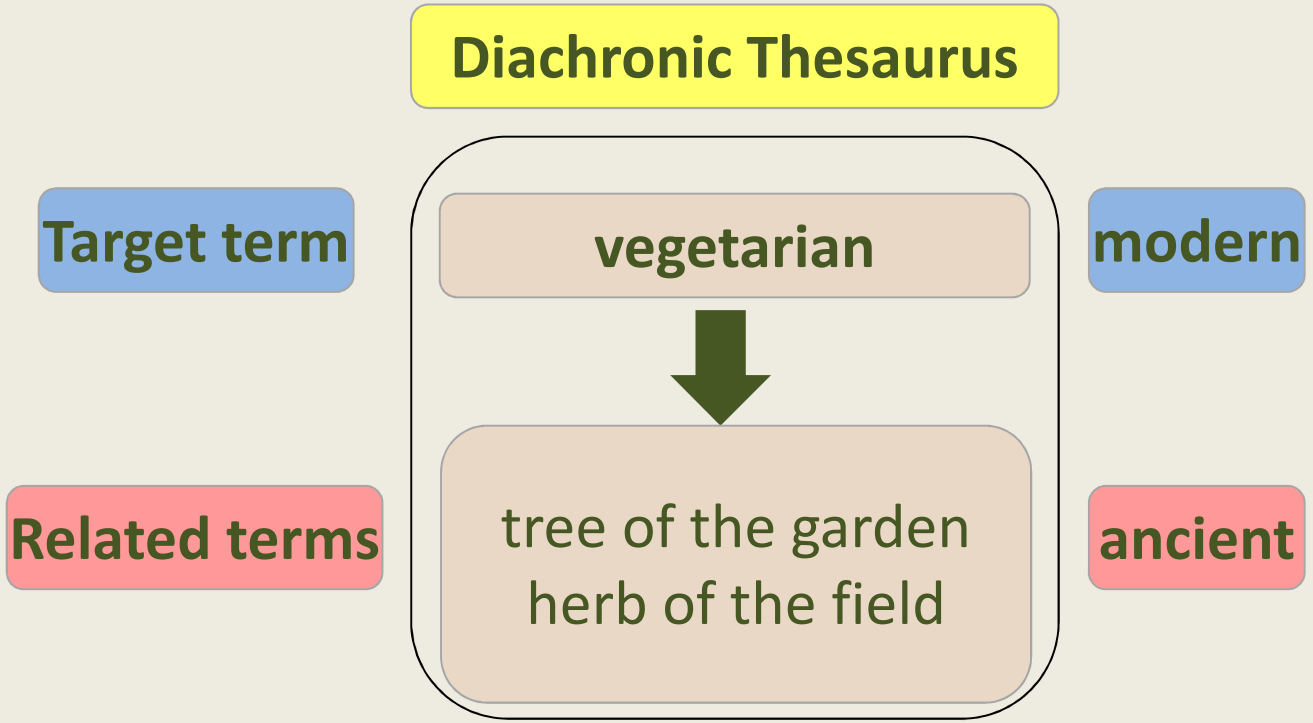
Of every tree of the garden thou mayest freely eat:
…
and thou shalt eat the herb of the field;

(King James Bible, Genesis)

Diachronic Corpus

Diachronic Thesaurus

A useful tool for supporting searches in diachronic corpus



Users are mostly aware of modern language

Diachronic Thesaurus

Prior work:
Collecting relevant related terms
- For given thesaurus entries

LaTeCH 2013

Our task:
Collecting a relevant list of modern target terms
- Domain/corpus dependent

Diachronic Thesaurus: Our Task

- Utilize a given candidate list of modern terms as input
- Predict which candidates are relevant for the domain corpus

✓ vegetarian
✓ ecology
× cell-phone
× computer

Background: Terminology Extraction (TE)

Corpus-based Terminology Extraction

1. ~~Automatically extract prominent terms from a given corpus~~
2. **Score candidate terms for domain relevancy**

Statistical measures for identifying prominent terms

Based on

- Frequencies in the target corpus (e.g. tf, tf-idf)

Or

- Comparison with frequencies in a reference background corpus

Supervised framework for TE

1. Candidate target terms are learning instances
2. Calculate a set of features for each candidate
3. Classification predicts which candidates are suitable

Baseline system (TE)

- Features : state-of-the-art TE scoring measures

Contributions

1. Integrating Query Performance Prediction in term scoring
2. Penetrating to ancient texts, via query expansion

Contribution #1

Integrating Query Performance Prediction in Term Scoring

Query Performance Prediction (QPP)

Estimate the retrieval quality of search queries

- **Assess quality of query results on the text collection.**

Our terminology scoring task

- QPP scoring measures are potentially useful – may capture additional aspects of term relevancy for the collection

TE		QPP
term is relevant for a domain	⇔	**term is a good query**

Query Performance Prediction (QPP)

Two types of statistical QPP methods

1. Pre-retrieval methods
   - Analyze query term's distribution within the corpus
2. Post-retrieval methods
   - Additionally analyze the top search results

Query Performance Prediction (QPP)

Integrate QPP measures as additional features

First integrated system (TE-QPP$_{\text{Term}}$**)**

- Applies the QPP measures to the candidate **term** as the query
- Utilizes these scores as additional classification features

Contribution #2

Penetrating to ancient texts

Penetrating to ancient periods

In a diachronic corpus

- A candidate term might be rare in its original modern form, yet frequently referred to by archaic forms

query term: vegetarian

Were All Men Vegetarians before the Flood?
…God instructed Adam saying, "I have given you every herb that yields…" (Genesis 1:29) …

(by Eric Lyons, M.Min.)

modern texts

Of every tree of the garden

every herb that yields

…

and thou shalt eat the herb of the field;

(King James Bible, Genesis)

ancient texts

Penetrating to ancient periods

Baseline (TE) and First integrated system (TE-QPP$_{Term}$)

- Rely on corpus occurrences of the original candidate term
  - Prioritize relatively frequent terms

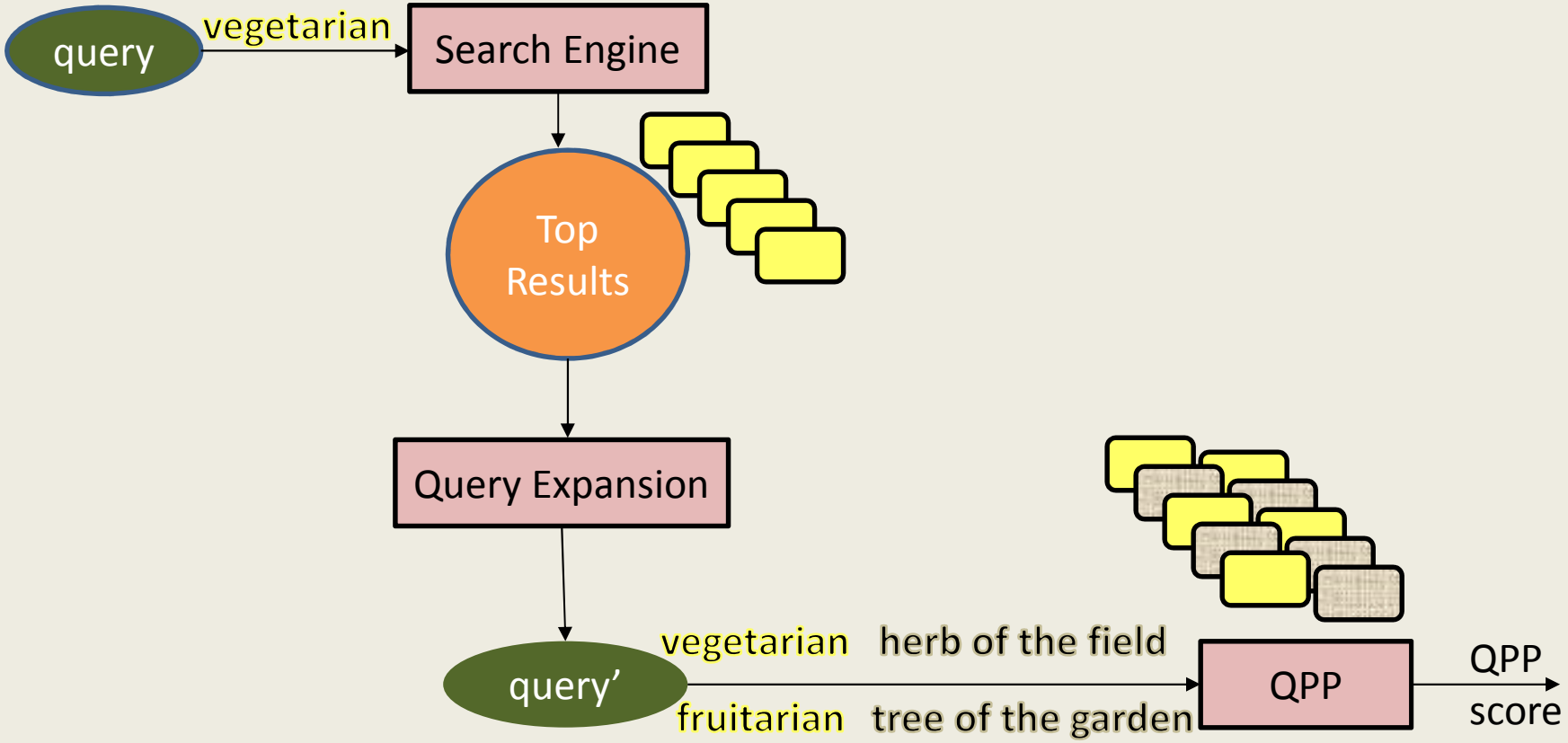
Our inspiration

- A post-retrieval QPP method
  - ✓ Query Feedback measure (Zhou and Croft, 2007)

Penetrating to ancient periods

Second integrated system (TE-QPP$_{QE}$**)**

- Utilizes Pseudo Relevance Feedback **Query Expansion**



Evaluation Setting

Diachronic corpus: the Responsa Project

✓Questions posed to rabbis along their detailed rabbinic answers
✓Written over a period of about a thousand years
✓76,760 articles
✓Used for previous IR and NLP research

Candidate target terms

- Hebrew Wikipedia entries
- Balanced for positive and negative examples
- #candidates: 500 train, 200 test

Classifier

Support Vector Machine with polynomial kernel

Results

Feature Set	Accuracy (%)
TE (baseline)	61.5
TE-QPP$_{Term}$	65
TE-QPP$_{QE}$	**66.5***

✓ Additional QPP features increase the classification accuracy

✓ Utilizing ancient documents, via query expansion, improves performance

✓ * Improvement over baseline statistically significant

- $p<0.05$ McNemar's test

Summary

Task: target term selection for a diachronic thesaurus

Main contributions:

1. Integrating Query Performance Prediction in Term Scoring
2. Penetrating to ancient texts via query expansion

Future work

- Utilize additional query expansion algorithms
  - Investigate the selective query expansion approach