Semantic Search Engine for Herbal Medicine

Team Herbert: Emmy, Gurdit, Ian, Karl
BUSINESS OVERVIEW
HERBAL AND ALTERNATIVE MEDICINE

- 1 billion health queries daily on Google, accounting for 7% of traffic
- Alternative medicine accounts for about 1/3 of health search topics.
- Herbal pharmaceuticals sales accounted for $8.8 billion in 2018 domestically and growth has been accelerating over the past decade.
- Where to look?
- Who to trust?
- How to align information on different websites?
Aggregate and distill the most reliable herbal medicine information into a curated single report view.

User who is curious about traditional Chinese medicine and finding credible information on its benefits and risks.
CURRENT SOLUTIONS

GENERAL PURPOSE SEARCH ENGINE

HERBAL MEDICINE WEBSITE

PROFESSIONAL MEDICAL WEBSITE

Too much data

Potentially low quality data

Overly technical data
OUR DIFFERENTIATIONS

**VERTICAL SEARCH ENGINE**
Herbert focuses on the search for semantic relationship between herbs and conditions.

**SEMANTIC RELATIONSHIP**
Herbert reliability extracts, identifies and develops semantic understandings of 3 main entities for each herbal medicine: herb, conditions and interactions with medications.

**MULTIPLE DATA SOURCES**
Herbert aggregates herbal medicine information from multiple data sources and prioritizes based on the number of data sources that reference them.

**LINKS TO ORIGINAL DATA**
Herbert provides links back to the original data source whenever possible for user to get more information.
DATA SOURCES

Scientific Journal of peer-reviewed biomedical literature

Chinese Herbal Medicine Database

MedlinePlus is an online health information service of the National Library of Medicine (NLM)

Integrative database of traditional Chinese medicine (TCM) focused on symptom mapping

Online encyclopedia where most general public use to find health information
MINIMUM VIABLE PRODUCT

2 PRIMARY USE CASES

- Search for information about a herb
- Search for herbs for a particular condition
DATA SCIENCE/TECHNICAL APPROACH
HERBERT AWS ARCHITECTURE

1) Data Mining Process
2) Summarization Process
3) Website

AWS Cloud

Database
RDS

Backend

Frontend

Bootstrap
jQuery
Jinja
Flask

S3 Bucket
EC2
EC2

web development, one drop at a time
HERBERT OVERALL ARCHITECTURE

Data Source
- Wikipedia
- PubMed
- MedlinePlus
- Me and Qi
- SymMap

(Step 1) Gather + Normalize
- Web Scrape, Data Extract & Transform
- Topic Selection
- Herb/Condition/Interaction Extraction
- Raw Normalized Content

(Step 2) Summarize
- Name Resolution
- Condition/Interaction Summarization
- Text Summarization
- Summarized Content

(Step 3) Query
- Web pages
- Search Engine
- Query Understanding
- Index Cache
**STEP 1: GATHER + NORMALIZE**

- **Data Source**
  - Wikipedia
  - PubMed
  - MedlinePlus
  - Me and Qi
  - SymMap

- **(Step 1) Gather + Normalize**
  - Web Scrape, Data Extract & Transform
  - Topic Selection
  - Herb/Condition/Interaction Extraction

- **Raw Normalized Content**

- **(Step 2) Summarize**
  - Name Resolution
  - Condition/Interaction Summarization
  - Text Summarization

- **Summarized Content**

**STEP 3: QUERY**

- **Search Engine**
  - Query Understanding
  - Index Cache

- **Web pages**
Step 1: Distill source data to relevant bullet points about an herb

Ginger is one of the first spices to have been exported from Asia, arriving in Europe with the spice trade, and was known to the ancient Greeks and Romans. The distantly related dicots in the genus *Asarum* are commonly called wild ginger despite their similar taste.
Step 1: Distill source data to relevant bullet points about an herb
Step 1: Distill source data to relevant bullet points about an herb

adverse, side-effect, toxic, interact, ...
Step 1: Distill source data to relevant bullet points about an herb

- adverse, side-effect, toxic, interact, ...
- safe, afet, fety
Step 1: Distill source data to relevant bullet points about an herb

- adverse, side-effect, toxic, interact, ...
- safe, afet, fety
Composition and safety

If consumed in reasonable quantities, ginger has few negative side effects,[37] it is on the FDA’s “generally recognized as safe” list,[38] though it does interact with some medications, including the anticoagulant drug warfarin[39] and the cardiovascular drug nifedipine.[2]

Chemistry

The characteristic fragrance and flavor of ginger result from volatile oils primarily consisting of zingerone, shogaols, and gingerols with [6]-gingerol and decanone as the major pungent compound.[40] Zingerone is produced by pungency and a spicy-sweet aroma.[40] Shagoals are more pungent and have higher antioxidant activity but not found in raw ginger, but is formed from gingerols during heating, storage or via acidity.[40]

Fresh ginger also contains an enzyme zingibain which is a cysteine protease and has similar properties to rennet.

Medicinal use and research

Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.[2][41][42][43] There is no clear evidence of harm from taking ginger during pregnancy, although its safety has not been established.[41][44] Ginger is not effective for treating dysmenorrhea,[45] and there is insufficient evidence for it having analgesic properties due to the lack of well conducted trials. Available data provides weak evidence for its anti-inflammatory role and it may reduce the subjective experience of pain in osteoarthritis.[46]

Allergic reactions to ginger generally result in a rash.[2] Although generally recognized as safe, ginger can cause heartburn and other side effects, particularly if taken in powdered form.[2] It may adversely affect individuals with gallstones and may interfere with the effects of anticoagulants, such as warfarin or aspirin.[2]
Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.\cite{1}\cite{2}\cite{3}\cite{4}\cite{5}\cite{6} There is no clear evidence of harm from taking ginger during pregnancy, although it may adversely affect individuals with gallstones and may interfere with the effects of anticoagulants, such as warfarin or aspirin.\cite{2}
Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent. [2][4][1][4][2][4][3] There is no clear evidence of harm from taking ginger during pregnancy, although
Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.\cite{41,42,43} There is no clear evidence of harm from taking ginger during pregnancy, although...
Health Benefits / Conditions

Nausea
CHALLENGES:
- Grabbing the relevant text
- No reliable single-step solutions
- Working without “labeled” data
- Working cross-domains (medical + cultural aspects)

LEARNINGS:
General, configurable pipeline that can be measured at each stage and used across multiple problems and domains.
(STEP 2) SUMMARIZE

Data Source
- Wikipedia
- PubMed
- MedlinePlus
- Me and Qi
- SymMap

(Step 1) Gather + Normalize
- Web Scrape, Data Extract & Transform
- Topic Selection
- Herb/Condition/Interaction Extraction

Raw Normalized Content

(Step 2) Summarize
- Name Resolution
- Condition/Interaction Summarization
- Text Summarization

Summarized Content

(Step 3) Query
- Web pages
- Search Engine
- Query Understanding
- Index Cache
Same herb can have many different names

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Me and Qi</td>
<td>Dittany root bark</td>
</tr>
<tr>
<td>SymMap</td>
<td>Dictamni Cortex</td>
</tr>
<tr>
<td>Wikibooks</td>
<td>Densefruit Pittany Root-Bark</td>
</tr>
<tr>
<td>Other names</td>
<td>White Fresh Bark</td>
</tr>
<tr>
<td>Other names</td>
<td>Burning Bush Root</td>
</tr>
<tr>
<td>Other names</td>
<td>Dictamnus dasycarpus root-bark</td>
</tr>
</tbody>
</table>

Chinese: 白鲜皮

Pinyin: Bái Xiān Pí
## CONDITION NORMALIZATION PROCESS

<table>
<thead>
<tr>
<th>Source</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Me and Qi</td>
<td>Dysmenorrhea</td>
</tr>
<tr>
<td>SymMap</td>
<td>Painful Periods</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Menstrual Cramps</td>
</tr>
<tr>
<td></td>
<td>Dysmenorrhea</td>
</tr>
</tbody>
</table>
George Shea (Kaiser Santa Clara Pharmacy Director for 13 pharmacies + TCM user): “This is immediately practical for the average layperson and overcomes the typical resources that are overly technical. Having cross references in itself is a major step forward in the field...I’ve seen many of these herbs in the TCM shops and am confident you hit the 80 in the 80-20 distribution of herbs.”
<table>
<thead>
<tr>
<th>Feedback</th>
<th>Example/Description</th>
<th>Team Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some of the medical names are too technical</td>
<td>E.g. hyperlipidemia vs. cholesterol</td>
<td>Manually mapped back some of technical names to “lay” names</td>
</tr>
<tr>
<td>Where can I find all the herbs/conditions?</td>
<td>Can only lookup herb through search or through hyperlinks, need prior knowledge</td>
<td>Added landing pages for master list of each</td>
</tr>
<tr>
<td>I can go from an herb to condition but not back</td>
<td>Hyperlinks on herb page but not conditions page</td>
<td>Added hyperlinks to conditions page</td>
</tr>
<tr>
<td>Looks like there are some typos in the summaries</td>
<td>E.g. Hanging empty parentheses, or new lines between sentences</td>
<td>Used regex to clean up some of the common “junk” patterns</td>
</tr>
</tbody>
</table>
STILL MORE TO DO

COOKING RECIPES

GINGER COOKBOOK
Healthy and Delicious Ginger Recipes

RESEARCH EVALUATION

DRUG INTERACTIONS AND NORMALIZATION

Navigating RxNorm Drugs
QUESTION & ANSWER
APPENDIX
RELEVANT TOPIC SELECTION EVALUATION

- **Metrics:** Precision, Recall, F1
  - Want to see how many relevant topics selected
  - Due to **high imbalance in content topics**, preferred over accuracy (average 2/6 are potentially condition-related)
  - Data: ‘50 fundamental herbs’
  - Agreement b/w annotators: **Fleiss Kappa of .65** (‘substantial’ but not perfect)

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>92%</td>
<td>63%</td>
<td>75%</td>
</tr>
</tbody>
</table>
CONDITION EXTRACTION EVALUATION

- Metrics: Precision, Recall, F1
- 75 entries across 5 Herbs
- Aimed to maximize precision for reliability

<table>
<thead>
<tr>
<th>Precision*</th>
<th>Recall*</th>
<th>F1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>94%</td>
<td>92% PA</td>
<td>68%</td>
</tr>
</tbody>
</table>

*Deflated/Pessimistic since we remove entries coming only from a single source that’s less trusted (e.g. Wikipedia only).
PA = “Post-adjustment”
<table>
<thead>
<tr>
<th>Source</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH</td>
<td>91%</td>
<td>48%</td>
<td>63%</td>
</tr>
<tr>
<td>Pubmed</td>
<td>100%</td>
<td>75%</td>
<td>86%</td>
</tr>
<tr>
<td>Wikipedia*</td>
<td>86%</td>
<td>80% PA</td>
<td>40%</td>
</tr>
<tr>
<td>Me &amp; Qi</td>
<td>97%</td>
<td>94%</td>
<td>96%</td>
</tr>
</tbody>
</table>

*Deflated/Pessimistic since we remove entries coming only from a single source that’s less trusted (e.g. Wikipedia only).
PA = “Post-adjustment”
Appendix: NAME ENTITY EXTRACTION EVALUATION

<table>
<thead>
<tr>
<th>Software Package</th>
<th>Processing Times Per Abstract (ms)</th>
<th>Processing Times Per Sentence (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP4J (java)</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Genia Tagger (c++)</td>
<td>73</td>
<td>3</td>
</tr>
<tr>
<td>BioNLP13CG</td>
<td>272</td>
<td>29</td>
</tr>
<tr>
<td>BioNLP13CG (TF + 12 CPUs)</td>
<td>72</td>
<td>7</td>
</tr>
<tr>
<td>jPDT (Dynet)</td>
<td>905</td>
<td>97</td>
</tr>
<tr>
<td>Dexter v2.1.0</td>
<td>208</td>
<td>84</td>
</tr>
<tr>
<td>MetaMapLite v3.6.2</td>
<td>293</td>
<td>89</td>
</tr>
<tr>
<td>en_core_sci_sm</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>en_core_sci_md</td>
<td>33</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>SOTA</th>
<th>+ Resources</th>
<th>sci sm</th>
<th>sci md</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC5CDR (Li et al., 2016)</td>
<td>83.87</td>
<td>86.92b</td>
<td>89.69bb</td>
<td>78.83</td>
<td>83.92</td>
</tr>
<tr>
<td>CLEF11 (Bau et al., 2011)</td>
<td>79.35</td>
<td></td>
<td></td>
<td>77.31</td>
<td>76.47</td>
</tr>
<tr>
<td>JNLPBA (Collier and Kim, 2004)</td>
<td>68.95</td>
<td>73.48b</td>
<td>75.50bb</td>
<td>71.78</td>
<td>73.21</td>
</tr>
<tr>
<td>BioNLP13CG (Pyysalo et al., 2015)</td>
<td>76.74</td>
<td></td>
<td></td>
<td>72.98</td>
<td>77.60</td>
</tr>
<tr>
<td>AnateEM (Pyysalo and Ananiadou, 2014)</td>
<td>88.55</td>
<td>91.61**</td>
<td></td>
<td>80.13</td>
<td>84.14</td>
</tr>
<tr>
<td>BC2GM (Smith et al., 2008)</td>
<td>84.41</td>
<td>80.51b</td>
<td>81.69bb</td>
<td>75.77</td>
<td>78.30</td>
</tr>
<tr>
<td>BC4CHEMD (Krallinger et al., 2015)</td>
<td>82.32</td>
<td>88.75a</td>
<td>89.37aa</td>
<td>82.24</td>
<td>84.55</td>
</tr>
<tr>
<td>Linnaeus (Gerner et al., 2009)</td>
<td>79.33</td>
<td>95.68**</td>
<td></td>
<td>79.20</td>
<td>81.74</td>
</tr>
<tr>
<td>NCBI-Disease (Dog an et al., 2014)</td>
<td>77.82</td>
<td>85.80b</td>
<td>87.34bb</td>
<td>79.50</td>
<td>81.65</td>
</tr>
</tbody>
</table>


Table 5: Test F1 Measure on NER for the small and medium scispacy models compared to a variety of strong baselines and state of the art models. The Baseline and SOTA (State of the Art) columns include only single models which do not use additional resources, such as language models, or additional sources of supervision, such as multi-task learning. + Resources allows any type of supervision or pretraining. All scispacy results are the mean of 5 random seeds.

- Compared to state of the art ("SOTA"), 2x the speed, 1/10th size, at the cost of about 7% difference in F1
TEXT SUMMARIZATION EVALUATION PLAN

● Metrics: Rouge-N Family
  ○ Like recall
  ○ Overlap of summary and “golden summary”

● Data: Plan to annotate and summarize a “train”, “test”, and “validation set”
  ○ Measure agreement b/w annotators

- ROUGE-N

\[
ROUGE-N = \frac{\sum_{s \in S} \sum_{g_n} C_m(g_n)}{\sum_{s \in S} \sum_{g_n} C(g_n)}
\]

where
- \( S \) is the set of manual summaries
- \( g_n \) is an individual manual summary
- \( C(g_n) \) is the number of co-occurrences of \( g_n \) in the manual summary and automatic summary
TEXT SUMMARIZATION EVALUATION

Need annotators to rate on qualitative scale since we have length-constrained extractive summary. Task strongly time intensive for humans.

Get measurement of agreement.

Not done yet due to time constraint.

For now at least from our own QA and SMEs, no major complaints...

Reference: “EVALUATION MEASURES FOR TEXT SUMMARIZATION” by Josef Steinberger & Karel Ježek
Appendix: TEXT SUMMARIZATION - TEXT RANK

Key Concept: Use Pagerank to score sentences with the assumption that “summary sentences” are similar and central to most other sentences.
Appendix: Website Index (Lookup Tags) V1

- Don’t want to comb every document every time we query so we build indices

- Key indices: Herbs, Conditions

- For each key index produce posting: (stemmed_word, [(documents containing word, length)], frequency stats)

- E.g. ‘ginger’ appear in 5 documents total

Posting:

('ging', [('Ginger', 150), ('Ginger Root', 165), ... ('Upset Stomach', 80)...], 100)
Appendix: Matching/Relevance Function V1

**Okapi Best Matching-25 (BM-25)**

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \left[ \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]$$

- **Document** (indices)
- **Query words** ($q_1 \ldots q_n$)
- **Reward rare words/Discount common words**
- **Discount long documents/Reward short documents**
- **Reward word matches**
- **Discount short documents that don't match at all**
Query: “Does ginger help with stomach pain?”

NER: ginger, help, stomach pain

Cosine Distance/Similarity

Index

NER

Thresholding

Website: Index and Matching V2
Appendix: RELEVANT TOPIC SELECTION

- Word vectors capture some meaning via distributional hypothesis.
- Incorporating character n-gram vectors can help capture structure.
- EX) n=4 for word “toxic”:
  <tox/toxi/oxic/xic>
  Where “<” denotes start, “>” denotes end.
- Now have way of making sense of out of vocab words like “toxicology”!
Appendix: NAME ENTITY EXTRACTION

- **Guts of model**: Hashed Word Embeddings + Parsing Features + Dilated n-gram CNN
- **Rough summary**: Look at meaning of words + grammatical grouping structure to determine if entity
- **Why not LSTM/CRF**: Speed of prediction, relative training size needed, relative simplicity of features
Appendix: Herb Name Resolution Process

Data Sources with English or Latin Herb Names
- MedlinePlus
  - Use herb names from MedlinePlus to look up herbs on Wikipedia and PubMed
- Wikipedia
- PubMed
  - 1) Extract English, Latin or other names
  - 2) Match Almost Exact names to find Pinyin name from TCM Reference Table

% of herbs found in ref. table
- MedlinePlus: 61%
- Wikipedia: 61%

Almost Exact means when the reference table herb name contains all the words (without stop words and no specific order) of the herb name

Data Sources with Pinyin Herb Names
- Fifty Fundamental Herbs
- Me and Qi
- Symmap
  - Match Exact Pinyin name with TCM Reference Table

% of herbs found in ref. table
- Fifty Fundamental Herbs: 74%
- Me and Qi: 94%
- Symmap: 86%

TCM Reference Table:
- Harvard TCM-ID Table has ~10000 TCM herb names and IDs
- Use BioBERT word embeddings to generate Pinyin name and find match from TCM Reference Table

% of herbs found in ref. table or among other data sources
- MedlinePlus: 65%*
- Wikipedia: 65%*
- Fifty Fundamental Herbs: 86%
- Me and Qi: 97%
- Symmap: 89%

*Note: MedlinePlus and Wikipedia are ~98% matched because we use MedlinePlus to guide Wikipedia herb search

Work-in-progress
Other potential Sticky facts to use (placeholder)

Health is the third most searched thing on internet
1. Email
2. Online Shopping
3. Healing

Google receives more than 1 billion health questions every day

Jackie Drees - Monday, March 11th, 2019 Print | Email

An estimated 7 percent of Google's daily searches are health-related, according to Google Health Vice President David Feinberg, MD, *The Telegraph* reports.
If We Could Do It Again...

- True to namesake: Use Bert Embeddings (Sequence to sequence rather than just words)
- Medical domain relation extraction system
Bring clarity and reduce search fatigue.