Incorporating WordNet in an Information Retrieval System

Project Advisor – Dr. Chris Pollett
Committee Members – Dr. Khuri and Dr. Mak
Presented by: Shailesh Padave
Agenda

• Introduction
• Query Expansion
• WordNet
• Part-of-Speech Tagging
• Similarity Ranking Functions
• Experiments and Conclusions
• Demo
Introduction

• Project Goal
  – Implement query expansion in Yioop
  – Extend query rewriting mechanism in Yioop to use Wordnet
  – Implement Part-Of-Speech tagging
  – Implement a Similarity Ranking Function
  – Rewrite a result reordering algorithm to use WordNet Scores
Query Expansion

• Reformulating a seed query to improve retrieval performance in information retrieval operations\cite{manning2008introduction}

• Different ways:
  – Finding synonyms of words Using WordNet
  – Techniques like spelling correction
  – Re-weighting the terms in the original query

• You would want a search for *computer*, then by query expansion we get
  – *Computing device*
  – *Information processing system*
  – *Data processor*

WordNet

• Founder – Dr. George A Miller, Princeton University\(^1\)
• Awarded the Antonio Zampolli Prize
• A Large Lexical Database for English or an “Electronic Dictionary”
• Covers English Verbs, Nouns, Adverbs, Adjectives
• Used in Many Information Retrieval Systems
• Useful tool for Computational linguistic and natural language processing
• Applications
  – Produce a combination of dictionary and thesaurus
  – Support automatic text analysis and artificial intelligence applications

WordNet

• Similar Applications
  – WordWeb, Artha, Moby thesaurus, openthesaurus etc.

• Large Database of English\(^1\)

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>Unique String</th>
<th>Synset</th>
<th>Total Word-Sense pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117798</td>
<td>82115</td>
<td>146312</td>
</tr>
<tr>
<td>Verb</td>
<td>11529</td>
<td>13767</td>
<td>25047</td>
</tr>
<tr>
<td>Adjective</td>
<td>21479</td>
<td>18156</td>
<td>30002</td>
</tr>
<tr>
<td>Adverb</td>
<td>4481</td>
<td>3621</td>
<td>5580</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>155287</strong></td>
<td><strong>117659</strong></td>
<td><strong>206941</strong></td>
</tr>
</tbody>
</table>
WordNet Database

• Database Information

<table>
<thead>
<tr>
<th>Type of word</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>data.adj, index.adj</td>
</tr>
<tr>
<td>Adverb</td>
<td>data.adv, index.adv</td>
</tr>
<tr>
<td>Noun</td>
<td>data.noun, index.noun</td>
</tr>
<tr>
<td>Verb</td>
<td>data.verb, index.verb</td>
</tr>
</tbody>
</table>

• Exceptions
  – noun.exec
  – verb.exec
  – adj.exec
  – adv.exec
Data.verb

- 00048819 29 v 01 habit 0 002 @ 00047662 v 0000 + 03479089 n 0101 01 + 09 00 | put a habit on

  - **synset_offset** - Current byte offset in the file (8 digit)
  - **lex_filenum** – (2 digit) lexicographer file name containing the synset
  - **ss_type** – n,v,a,r
  - **w_cnt** – number of words in synset (2 digit HEX)
  - **word** – Actual search word
  - **Lex_id** – a hexadecimal digit appended to lexicographic file
  - **P_cnt** – count of pointers
    - **Pointer_symbol** – define a relationship with other words
    - **Synset_offset**
    - **Part of speech**
    - **First 2 HEX digits for source and next 2 digits for target**
  - **gloss** – represented as vertical bar followed by text string. May contain 1 or more examples
Index.verb and noun.exec

- **body v 1 2 @ 1 0 02672913**
  - lemma – lower case ASCII text of the word
  - pos – n v a r (part of speech)
  - synset_cnt – number of synsets that lemma is in
  - p_cnt – number of pointers
  - Pointer_symbol - @ for hypernym, ! For antonyms, etc. otherwise p_cnt is 0
  - Sense_count – number of senses
  - Tagsense-count – number of tags
  - Synset_offset – 8 digit offset used in data.pos

- **corpora corpus**
  - Irregular word
  - Base form of word
Output of WordNet

• Input word - fly

The noun fly has 5 senses (first 4 from tagged texts)

1. (6) fly -- (two-winged insects characterized by active flight)
2. (1) tent-fly, rainfly, fly sheet, fly, tent flap -- (flap consisting of a piece of canvas that can be drawn back to provide entrance to a tent)
3. (1) fly, fly front -- (an opening in a garment that is closed by a zipper or by buttons concealed under a fold of cloth)
4. (1) fly, fly ball -- ((baseball) a hit that flies up in the air)
5. fly -- (fisherman's lure consisting of a fishhook decorated to look like an insect)

The verb fly has 14 senses (first 9 from tagged texts)

1. (33) fly, wing -- (travel through the air; be airborne; "Man cannot fly")
2. (9) fly -- (move quickly or suddenly; "He flew about the place")
3. (5) fly, aviate, pilot -- (fly a plane)
4. (3) fly -- (transport by aeroplane; "We fly flowers from the Caribbean to North America")
5. (2) fly -- (cause to fly or float; "fly a kite")
6. (2) fly -- (be dispersed or disseminated; "Rumors and accusations are flying")
7. (2) fly -- (change quickly from one emotional state to another; "fly into a rage")
Output From Command Line

- Input word – *fly (wn <search_word> -over)*

**Overview of noun fly**

The noun fly has 5 senses (first 4 from tagged texts)

1. (6) fly — (two-winged insects characterized by active flight)
2. (1) tent-fly, rainfly, fly sheet, fly, tent flap — (flap consisting of a piece of canvas that can be drawn back to provide entrance to a tent)
3. (1) fly, fly front — (an opening in a garment that is closed by a zipper or by buttons concealed under a fold of cloth)
4. (1) fly, fly ball — (baseball) a hit that flies up in the air
5. fly — (fisherman’s lure consisting of a fishhook decorated to look like an insect)

**Overview of verb fly**

The verb fly has 14 senses (first 9 from tagged texts)

1. (33) fly, wing — (travel through the air; be airborne; "Man cannot fly")
2. (9) fly — (move quickly or suddenly; "He flew about the place")
3. (5) fly, aviate, pilot — (fly a plane)
4. (3) fly — (transport by aeroplane; "We fly flowers from the Caribbean to North America")
Query Expansion Example

- Consider an input query: computing device

<table>
<thead>
<tr>
<th>Synonyms for computing</th>
<th>Synonyms for device</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer science</td>
<td>gimmick</td>
</tr>
<tr>
<td>calculation</td>
<td>twist</td>
</tr>
<tr>
<td>computation</td>
<td>device</td>
</tr>
<tr>
<td>cipher</td>
<td></td>
</tr>
<tr>
<td>work out</td>
<td></td>
</tr>
<tr>
<td>reckon</td>
<td></td>
</tr>
<tr>
<td>calculate</td>
<td></td>
</tr>
</tbody>
</table>
Query Expansion Example

- Total possible combinations for *computing device* – 21

<table>
<thead>
<tr>
<th>Combinations for computing device</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>computer science gimmick</td>
<td>computer science twist</td>
<td>computer science device</td>
<td>calculation gimmick</td>
</tr>
<tr>
<td>calculation twist</td>
<td>calculation device</td>
<td>computation gimmick</td>
<td>computation twist</td>
</tr>
<tr>
<td>computation device</td>
<td>cipher gimmick</td>
<td>cipher twist</td>
<td>cipher device</td>
</tr>
<tr>
<td>work out gimmick</td>
<td>work out twist</td>
<td>work out device</td>
<td>reckon gimmick</td>
</tr>
<tr>
<td>reckon twist</td>
<td>reckon device</td>
<td>calculate gimmick</td>
<td>calculate twist</td>
</tr>
<tr>
<td>calculate device</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Part Of Speech Tagging

• Process of marking up or tagging a word based on definition and context
• Also called as POS tagging / POST
• Decided by its relationship with adjacent word
• Two approaches
  – Rule based
  – stochastic
Part-of-Speech Tagging continued..

• Rule based POST is Oldest approach
• Use of Hand Written Rules and Dictionary
• Difficult to automate the process
• Easy way to use corpus containing manual tagging and linguistic rules
• Bigger lexicon, more processing time
Part-of-Speech Tagging continued..

• Word ends with ‘ed’ is past participate, ‘ly’ is adverb
• Use of Brill tagger
• Common Tags are
  – NN - Singular or mass Noun
  – NNS - Plural Noun
  – JJ – Adjective
  – IN- Preposition
• Use of Linguistic feature of word
  – We~NN systematically~AV analyze~NN the~DT performance~NN of~IN these~DT techniques~NN versus~IN existing~VB search~NN results~NN
Implementation of Part-Of-Speech Tagger for WordNet

• Placed between Yioop and WordNet Layer

• Experimented Part-of-Speech Tagging
  – During Crawl Time
  – During Query Time
Part Of Speech Tagging

• Consider input – running dog
• After Part-Of-Speech Tagging
  – running~\textsc{VB} dog~\textsc{NN}
• Get WordNet Result for word \textit{running}
  – Noun has 5 senses
  – Verb has 41 senses
  – Adjective has 7 senses
• Total we have 52 senses
Part Of Speech Tagging

- For running, VB is part-of-speech.
- Get verb senses from WordNet search result

The verb run has 41 senses (first 29 from tagged texts)

1. (106) run -- (move fast by using one's feet, with one foot off the ground at any given time; "Don't run--you'll be out of breath"; "The children ran to the store")
2. (38) scat, run, scap, turn tail, lam, run away, high tail it, bunk, head for the hills, take to the woods, escape, fly the coop, break away -- (flee; take to one's heels; cut and run; "If you see this man, run!"; "The burglars escaped before the police showed up")
3. (21) run, go, pass, lead, extend -- (stretch out over a distance, space, time, or scope; run or extend between two points or beyond a certain point; "Service runs all the way to Cranbury"; "His knowledge doesn't go very far"; "My memory extends back to my fourth year of life"; "The facts extend beyond a consideration of her personal assets")
4. (20) operate, run -- (direct or control; projects, businesses, etc.; "She is running a relief operation in the Sudan")

- Out of 52, we will work on 41 senses
- Improved Processing, execution Speed
How to Extract Similar Words?

• We have Similar words, exact meaning and usage of similar words in sentence
  
  (106) run -- (move fast by using one's feet, with one foot off the ground at any given time; "Don't run you'll be out of breath"; "The children ran to the store")

• Use of Similarity Ranking Functions
• Methods to find similarity between two sentences
• We used Cosine Similarity Ranking, Intersection Ranking, Okapi BM25 ranking
Cosine Similarity Ranking

• Measure of Similarity between two vectors of an inner space product
• Independent of magnitude of vectors
• Should be in positive space
• Term Frequency (TF)

\[ TF = \log(f_{t,d}) + 1 \text{ if } f_{t,d} > 0 \text{ & } 0 \text{ otherwise} \]

• IDF (Inverse Document Frequency)

\[ IDF = \log\left(\frac{N}{N_t}\right) \]

Cosine Similarity Ranking

• Given two \(|V|\) dimensional vectors as \(x\), for query and \(y\) for document
  
  \[ \hat{x} = (x_1, x_2, x_3, x_4, \ldots, x_{|v|}) \]
  
  \[ \hat{y} = (y_1, y_2, y_3, y_4, \ldots, y_{|v|}) \]

• Dot product of \(x\) and \(y\) is given as
  
  \[ \hat{x} \cdot \hat{y} = \sum_{i=1}^{|v|} x_i \cdot y_i \]
Cosine Similarity Ranking

- Geometric meaning is
  \[ \hat{x} \cdot \hat{y} = |\hat{x}| |\hat{y}| \cos \theta \]

- The length of the vector \(|\vec{v}| = \sqrt{\sum_{i=1}^{|V|} v_i^2}\)

- To calculate the angle
  \[ \cos \theta = \frac{\sum_{i=1}^{V} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{V} x^2} \sqrt{\sum_{i=1}^{V} y^2}} \]

- Two vectors are collinear if \( \theta = 0^\circ, \cos \theta = 1 \)
- Two vectors are orthogonal if \( \theta = 90^\circ, \cos \theta = 0 \)
Cosine Similarity Ranking

- Consider a query vector $\vec{q}$ and document vector $\vec{d}$, then the similarity is defined as the cosine of the angle between them

$$sim(\vec{d}, \vec{q}) = \frac{\vec{d} \cdot \vec{q}}{|\vec{d}| \cdot |\vec{q}|}$$
Intersection Ranking

• Split both sentences into array of words known as *tokens*
• Get common tokens between two sentences
• Intersection Ranking computed as follows:

\[
f(s_1, s_2) = \frac{|\{w | w \text{ in } s_1 \& w \text{ in } s_2\}|}{(|s_1| + |s_2|)/2}
\]

\(|s_1| \text{ and } |s_2| \text{ is the length of documents } s_1 \text{ and } s_2 \text{ respectively}\)
Index Manager in Yioop

• Contains inverted index
• Provides mapping between terms and their locations
• Two main components
  – Dictionary – terms in the vocabulary
  – Posting list – position of term in collection
• Example:
  – Device -> (1,2207), (20,4678), .... ... , (22,127838)
  – Engineering -> (2,36374), (9,667778)
# Counts from Index Manager

<table>
<thead>
<tr>
<th>Expanded Query</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer science device</td>
<td>43</td>
</tr>
<tr>
<td>Cipher device</td>
<td>32</td>
</tr>
<tr>
<td>Work out device</td>
<td>30</td>
</tr>
<tr>
<td>Computational device</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Okapi BM25

• Retrieval function for Bag-of-words
• Rank a set of documents depending upon appearance of query terms in each document
• Independent of relative proximity of query words in document
• Two Important factors: $TF, IDF$
Okapi BM25

- Term Frequency (TF) is calculated as

\[
TF_{BM25} = \frac{f_{t,d} \cdot (k_1 + 1)}{f_{t,d} + k_1 \cdot \left( (1 - b) + b \cdot \left( \frac{l_d}{l_{avg}} \right) \right)}
\]

- The Inverse Document Frequency (IDF) is calculated as

\[
IDF(t) = \log \left( \frac{N}{N_t} \right)
\]

- The BM25 scoring function is defined as

\[
Score_{BM25}(q, d) = \sum_{t \in q} IDF(t) \cdot TF_{BM25}(t, d)
\]
Usage of Similarity Ranking Algorithms

- To extract exact similar words from WordNet
- To sort search results after query expansion
Design & Implementation

Yioop → Part of speech tagger → WordNet

Reordering of web pages according to WordNet Score → BM25 Score Calculation

WordNet result extraction → Similarity Ranking function

Index Manager
Sequence Diagram

Integration of WordNet in Yioop Search Engine

- Yioop
- Part-Of-Speech Tagger
- WordNet
- Extractor
- Similarity Function
- Index Manager
- WordNet Score Calculator
- BM25 Score Calculator

1. Yioop sends an Input Query
2. Yioop requests for the Search result
3. Similar Words Extraction Request
4. Request for Similarity Score
5. Request for Ordered Words
6. Request for Count from Index
7. Request for Top two similar words
8. Request for BM25 score
9. Sends WordNet score

alt [if WordNet Scores is greater zero]

- Reorder as per BM25 score

alt [if WordNet Scores is zero]

- Order using RRF score
Experiments

• Used three datasets as follows:

<table>
<thead>
<tr>
<th>Name of Dataset</th>
<th>Number of crawled pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJSU CS</td>
<td>18156</td>
</tr>
<tr>
<td>Wikipedia dataset</td>
<td>100,000</td>
</tr>
<tr>
<td>dmoz dataset</td>
<td>972800</td>
</tr>
</tbody>
</table>

• Effectiveness of retrieved method is measured by the relevance provided by human assessment

• Two important aspects: Recall & Precision
Recall and Precision

• Recall – the fraction of relevant documents which appears in result set
• Precision - the fraction of result set which is relevant
• Experimented Part Of Speech Tagging
  – During Crawl time and Query Time
  – During Query Time
Experimentation with Part Of Speech Tagging During Crawl

Without Part-Of-Speech Tagging

With Part-Of-Speech Tagging
Experiment on Part Of Speech Tagging
Screen Shots

• WordNet Results

![WordNet Results: computing device, computing machine](image)

• WordNet Score

<table>
<thead>
<tr>
<th>WordNet Results: luggage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compagnia del Viaggio - Top Business Consumer_Goods_and_Services</td>
</tr>
<tr>
<td>Luggage_and_Bags Bags_and_Back_Pac</td>
</tr>
<tr>
<td><a href="http://www.compagniadiviaggio.it">www.compagniadiviaggio.it</a></td>
</tr>
<tr>
<td>Italy. Suitcases, travel bags, beauty cases, hand baggage and toiletry cases.</td>
</tr>
<tr>
<td>Cached, Similar, Inlinks, IP: 205.188.95.207, Score: 9.84</td>
</tr>
</tbody>
</table>

| Charlotte of America - Top Business Transportation and Logistics Aviation |
| Ground_Support_Equipment |
| www.charlotte.com |
| Manufacturer of ground support equipment, including baggage tractors and mobile baggage loaders. |
| Cached, Similar, Inlinks, IP: 64.12.249.187, Score: 10.0 |
Important Findings

• Recall increases, precision decreases and vice-versa

• Number of URLs visited
  – Using Part-of-speech tagging during crawl time = 660 / hour
  – Not Using Part-of-speech tagging during crawl time = 800 / hour
Important Findings

• In absence of WordNet score, search results will sort by RRF (Reciprocal Rank Fusion)
• RRF is addition of Doc Rank score, Relevance score and Proximity score after normalization
• Sorted search results are more relevant as per order.
• Top two WordNet words are shown on Search Result window.
Comparison with WordNet score and RRF

Reciprocal Rank for Query - information technology

WordNet Score for Query - information technology
<table>
<thead>
<tr>
<th>Website</th>
<th>WordNet Rank</th>
<th>RRF Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligroup, Inc. - Top Business Information_Technology Consulting</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>David Christie and Associates Pty Ltd - Top Business Information_TechnologyEmployment Recruitment_</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>The Campus Computing Project - Top Reference Education Colleges_and_Universities North_America Unit</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>The International Foundation for Information Technology (IF4IT) - Top BusinessInformation_Technolo</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Information Technology and Business Teachers of Ireland - Top Regional Europe Ireland Education</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>N. Voustros Information Technology Services - Top Regional Europe Greece Prefectures Ioannina Busin</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Recruiter Solutions International - Top Business Employment Recruitment_and_Staffing Recruiters</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Estonian Association of Information Technology and Telecommunications - ITL - Top Regional Europe</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>International Federation for Information Technology and Tourism</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>ADEC Distance Education Consortium</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>
Conclusion

- Query expansion helped to improve search results.
- Part-of-speech tagger helps efficient implementation.
- Similarity Ranking algorithms helped to retrieve exact words from WordNet.
- WordNet feature works for Windows, Linux, Mac.
- Part of speech tagging during crawl time is not efficient, so used only during crawl time.
- Throughput time is increased by 0.2 seconds and deviation of 0.3 seconds.
- WordNet works only for English language.
Future Scope

• Yioop is multi-language search engine
• Implement query expansion with other languages as well.
• WordNet Feature is flexible enough to adopt new English dictionary.
• WordNet has many other feature, we can use one of them as improvement in search engine feature.
References


Thank You