Semantic Similarity in a Taxonomy

By
Ankit Ramteke (113050067)
Bibek Behera (113050043)
Karan Chawla (113050053)
• It is based on a paper

“Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language”

By Philip Resnik (July 1999)
Outline

1) Motivation
2) Abstract
3) Similarity Measures and Implementation
4) Resolving Syntactic Ambiguity using Semantic Similarity Measure
5) Using taxonomic similarity in word sense selection
6) Conclusions
7) References
Motivation

- Evaluating semantic relatedness using network representations is a difficult problem
- One Approach: Edge Counting method
- For example:
Motivation Contd.

• **Drawback with Edge-Counting Approach:**
  It relies on the notion that links in the taxonomy represent uniform distances.
Example

Example:

- Rabbit Ears
  Sense 1
  rabbit ears -- (an indoor TV antenna; consists of two extendible rods that form a V)
    => television antenna, tv-antenna -- (an omnidirectional antenna tuned to the broadcast frequencies assigned to television)

- phytoplankton
  Sense 1
  phytoplankton -- (photosynthetic or plant constituent of plankton; mainly unicellular algae)
    => plant, flora, plant life -- (a living organism lacking the power of locomotion)
      => organism, being -- (a living thing that has (or can develop) the ability to act or function independently)
        => living thing, animate thing -- (a living (or once living) entity)
Abstract

- This paper presents a semantic similarity measure in an IS-A taxonomy.
- It presents algorithms that take advantage of taxonomic similarity in resolving syntactic and semantic ambiguity.
Some Notations

- $C$ : Set of concepts in an is-a taxonomy
- $p(C) :$ probability of encountering an instance of concept $c$
  
  if $c_1$ is-a $c_2$, then $p(c_1) \leq p(c_2)$

- Information content of a concept $c = - \log p(c)$
Similarity Measure

- Similarity between concepts $c_1$ and $c_2$,
  \[
  \text{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)]
  \]
  Where,
  - $S(c_1, c_2)$: set of concepts that subsume both $c_1$ and $c_2$

- Similarity between words $c_1$ and $c_2$,
  \[
  \text{wsm}(w_1, w_2) = \max_{c_1, c_2} [\text{sim}(c_1, c_2)]
  \]
  where,
  - $c_1$ ranges over $s(w_1)$
  - $c_2$ ranges over $s(w_2)$
Concept and Word Similarity

<table>
<thead>
<tr>
<th></th>
<th>c1(description)</th>
<th>c2(description)</th>
<th>Subsumer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor1</td>
<td>(medical)</td>
<td>Nurse1</td>
<td>Health_Professional</td>
</tr>
<tr>
<td>(medical)</td>
<td>(medical)</td>
<td></td>
<td>Person</td>
</tr>
<tr>
<td>Doctor2</td>
<td>(PH D.)</td>
<td>Nurse2</td>
<td>Person</td>
</tr>
<tr>
<td>(PH D.)</td>
<td>(medical)</td>
<td></td>
<td>Person</td>
</tr>
</tbody>
</table>
Implementation

- The concept probability is calculated as
  \[ p(c) = \frac{\text{freq}(c)}{N} \]

Where,

- \( \text{freq}(c) = \sum_{n \in \text{words}(c)} \text{count}(n) \)
- \( N : \text{Total Number of nouns observed} \)
Other Measures

2) Edge-counting method

\[ \text{wsim}_{\text{edge}}(w_1, w_2) = (2 \times \text{MAX}) - [ \min_{c_1, c_2} \text{len}(c_1, c_2) ] \]

3) Probability method

\[ \text{sim}_{p(c)}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [1 - p(c)] \]

\[ \text{wsim}_{p(c)}(w_1, w_2) = \max_{c_1, c_2} [ \text{sim}_{p(c)}(c_1, c_2) ] \]
Results

<table>
<thead>
<tr>
<th>Similarity Method</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Method (replication)</td>
<td>r = 0.9015</td>
</tr>
<tr>
<td>Information Content</td>
<td>r = 0.7911</td>
</tr>
<tr>
<td>Probability</td>
<td>r = 0.6671</td>
</tr>
<tr>
<td>Edge-Counting</td>
<td>r = 0.6645</td>
</tr>
</tbody>
</table>

r : Correlation between the similarity ratings and the mean ratings reported by Miller and Charles.
Critique

- The measure sometimes produces spuriously high similarity measures for words on the basis of inappropriate word senses.
- For example:

<table>
<thead>
<tr>
<th>n1</th>
<th>n2</th>
<th>wsim(n1,n2)</th>
<th>subsumer</th>
</tr>
</thead>
<tbody>
<tr>
<td>tobacco</td>
<td>alcohol</td>
<td>7.63</td>
<td>Drug</td>
</tr>
<tr>
<td>tobacco</td>
<td>sugar</td>
<td>3.56</td>
<td>Substance</td>
</tr>
<tr>
<td>tobacco</td>
<td>horse</td>
<td>8.26</td>
<td>Narcotic</td>
</tr>
</tbody>
</table>
Solution

• Take a weighted similarity measure:

\[ \text{wsim}_\alpha(w_1, w_2) = \sum_i \alpha(c_i) [- \log p(c_i)] \]

where,

\[ \alpha(c_i) : \text{weight for each sense} \]
Using Taxonomic Similarity in Resolving Syntactic Ambiguity
Motivation

• “every way ambiguous” syntactic constructions
• Prepositional phrases, nominal compounds
  – I saw a boy with a telescope
  – A bank and a warehouse guard
• This section investigates role of semantic similarity in disambiguating nominal compounds
Motivation

- Noun phrase co-ordinations of the form *n1 and n2 n3*
- There are two possibilities:
  - A (bank and warehouse) guard
  - A (policeman) and (park guard)
- It is important to distinguish between the two
Why semantic similarity??

• Similarity of form and similarity of meaning are important cues for conjoinability

• Similarity of form can be captured by agreement in number (singular vs. plural)
  – Several business and university groups
  – Several businesses and university groups

• Boolean variable : candidate heads satisfying number agreement
Why semantic similarity??

• Similarity of meaning of conjoined heads
  – A television and radio personality
  – An actor and stage performer

• Television and radio are more similar in meaning so are actor and performer

• This similarity is captured well by semantic similarity in a taxonomy

• Thus the second measure would be semantic similarity using information content
Some more examples

• Noun-Noun modification
  – Mail and securities fraud
  – Corn and peanut butter
• Mail fraud is an important nominal phrase
• Corn butter??
• Now consider
  – Corn and peanut crops
Selectional Association

- Lexical co-occurrence and semantic class membership

- Selectional association of word \( w \) with a WordNet class \( c \) is given by

\[
-A(w, c) = \frac{p(c|w) \log \frac{p(c|w)}{p(c)}}{D(p(C|w) || p(C))}
\]

Here \( D(p1||p2) \) is the Kullback-Liebler distance between probability distributions \( p1 \) and \( p2 \). It is basically used for normalization.
Selectional Association

• For example, $A(\text{wool, clothing})$ would have a higher value than, say, $A(\text{wool, person})$.

• $A(w1; w2)$ of two words is defined as the maximum of $A(w1; c)$ taken over all classes $c$ to which $w2$ belongs.

• For example, $A(\text{wool, glove})$ would most likely be equal to $A(\text{wool, clothing})$, as compared to, say, $A(\text{wool, sports equipment})$
Using Selectional Association

• Resolving coordination ambiguities as follows:
  • Prefer analyses that strong affinities
    – Bank as a modifier of guard
  • Dis-prefer very weak relationships
    – Corn as a modifier of butter
• Strong and weak would be parameters of the algorithm
Experiments

• Two sets of 100 noun phrases
  – Form [NP n1 and n2 n3]
  – Wall Street Journal (WSJ) corpus, as found in the Penn Treebank.
• Disambiguated manually
• One set used for training and other for testing
• Preprocessing
  – All names were replaced with “someone”
  – Expansion of month abbreviations
  – Reduction of nouns to their root forms
Calculation of Measures of similarity

• Number Agreement
  – Simple analysis of suffixes with WordNet’s lists of root nouns and irregular plurals

• Semantic Similarity
  – \(w_{sim}(w_1, w_2) = \max_{c_1, c_2}[\text{sim}(c_1, c_2)]\)
  – Sample of approximately 800,000 noun occurrences in Associated Press Newswire stories used for probabilities calculation
  – Nouns not in WordNet were treated as belonging to class “thing”
Calculation of Measures of similarity

• Selectional association
  – Co-occurrence frequencies taken from a sample of approximately 15,000 noun-noun compounds extracted from the WSJ corpuses
## Decision Rules

<table>
<thead>
<tr>
<th>Source of evidence</th>
<th>Conjoined</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number agreement</td>
<td>n1 and n2</td>
<td>number(n1) = number(n2) AND number(n1) ≠ number(n3)</td>
</tr>
<tr>
<td></td>
<td>n1 and n3</td>
<td>number(n1) = number(n3) AND number(n1) ≠ number(n2)</td>
</tr>
<tr>
<td></td>
<td>undecided</td>
<td>otherwise</td>
</tr>
<tr>
<td>Semantic similarity</td>
<td>n1 and n2</td>
<td>wsim(n1,n2) &gt; wsim(n1,n3)</td>
</tr>
<tr>
<td></td>
<td>n1 and n3</td>
<td>wsim(n1,n3) &gt; wsim(n1,n2)</td>
</tr>
<tr>
<td></td>
<td>undecided</td>
<td>otherwise</td>
</tr>
<tr>
<td>Noun-noun modification</td>
<td>n1 and n2</td>
<td>A(n1,n3) &gt; (\tau) OR A(n3,n1) &gt; (\tau)</td>
</tr>
<tr>
<td></td>
<td>n1 and n3</td>
<td>A(n1,n3) &lt; (\sigma) OR A(n3,n1) &lt; (\sigma)</td>
</tr>
<tr>
<td></td>
<td>undecided</td>
<td>otherwise</td>
</tr>
</tbody>
</table>
Combination of methods

• (a) A simple form of “backing off”
  – Number agreement
  – Selectional association
  – Semantic similarity
• (b) Taking a vote among the three strategies and choosing the majority
• (c) Classification using the results of a linear regression
• (d) construction of a decision tree classifier
<table>
<thead>
<tr>
<th>Strategy</th>
<th>Coverage (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>100.0</td>
<td>66.0</td>
</tr>
<tr>
<td>Number agreement</td>
<td>53.0</td>
<td>90.6</td>
</tr>
<tr>
<td>Noun-noun modification</td>
<td>75.0</td>
<td>69.3</td>
</tr>
<tr>
<td>Semantic similarity</td>
<td>66.0</td>
<td>71.2</td>
</tr>
<tr>
<td>Backing off</td>
<td>95.0</td>
<td>81.1</td>
</tr>
<tr>
<td>Voting</td>
<td>89.0</td>
<td>78.7</td>
</tr>
<tr>
<td>Number agreement + DEFAULT</td>
<td>100.0</td>
<td>82.0</td>
</tr>
<tr>
<td>Noun-noun modification + DEFAULT</td>
<td>100.0</td>
<td>65.0</td>
</tr>
<tr>
<td>Semantic similarity + DEFAULT</td>
<td>100.0</td>
<td>72.0</td>
</tr>
<tr>
<td>Backing off + DEFAULT</td>
<td>100.0</td>
<td>81.0</td>
</tr>
<tr>
<td>Voting + DEFAULT</td>
<td>100.0</td>
<td>76.0</td>
</tr>
<tr>
<td>Regression</td>
<td>100.0</td>
<td>79.0</td>
</tr>
<tr>
<td>ID3 Tree</td>
<td>100.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>
Results

• Voting
  – A choice was made in 89 cases.
  – The number agreement agreed in 57 cases, of which 43 (75.4%) were correct
  – The semantic similarity strategy agreed in 58 cases, of which 43 (74.1%) were correct
  – The Selectional association strategy agreed in 73 cases, of which 50 (68.5%) were correct
Results

• Backing Off
  – Number agreement makes a choice for 53 cases and is correct for 48 (90.6%)
  – Selectional association makes a choice for 35 cases and is correct for 24 (68.6%)
  – Semantic similarity makes a choice for 7 cases of which 5 are correct (71.4%)
  – The remaining 5 cases are undecided
Comments

• Semantic similarity does contain information about the correct answer:
  – Agrees with the majority vote a substantial portion of the time (74%)
  – Selects correct answers more often than one would expect by default for the cases it receives through backing off
Using taxonomic similarity in word sense selection
Using taxonomic similarity in word sense selection

• Selecting the sense of noun when they appear in context of other nouns similar in meaning.

• Query expansion using related words is a well studied technique in information retrieval.

• Clusters of similar words can play a role in smoothing stochastic language models for speech recognition.

• Clusterings of related words can be used in characterizing subgroupings of retrieved documents in large-scale Web searches.
Assosciating word senses with noun groupings

For some tasks, the relevant relationships are not among words but word sense.
Using this cluster if we search, we will retrieve documents involving advice (one sense of counsel) and royalty (one sense of count)
Resnik's algorithm

- Resnik introduced an algorithm in 1998 which finds the most relevant group of word sense based on relationship between pairs of word in a group.

- The algorithm is as follows:

  Let the set of words be

  \[ W = \{ w_1, w_2 \ldots w_n \} \]

  - Each word \( w_i \) has a sense set

  \[ s_i = \{ s_{i,1}, s_{i,2} \ldots s_{i,m} \} \]
• For example: -

\[ W = \{ \text{doctor, nurse, actor} \} \]

\[ s_{\text{doctor}} = \{ \text{medical doctor, phd holder} \} \]

\[ s_{\text{nurse}} = \{ \text{health professional, nanny} \} \]

\[ s_{\text{actor}} = \{ \text{a theatrical performer} \} \]
• For each word pair, we find the nearest ancestor($c_{1,2}$) and the word similarity value($v_{1,2}$).

For example:

\[ w_1 = \text{doctor} \quad w_2 = \text{nurse} \]
\[ c_{1,2} = \text{health professional} \]
\[ v_{1,2} = 8.844 \]

Now we traverse through all the senses for each word $w_i$. The idea is that a sense is more relevant if it is a child of the ancestor of the pair of words. Such an ancestor gives support to that sense by incrementing its support value by its information content.
\[
\text{support (medical doctor)} = \text{support (medical doctor)} + v_{1,2}
\]

\[
\text{support (medical doctor)} = \text{support (medical doctor)} + 8.844
\]

\[
\text{support (health professional)} = \text{support (health professional)} + 8.844
\]

\[
\text{support (phd)} = \text{support (phd)} + 0
\]

\[
\text{support (nanny)} = \text{support (nanny)} + 0
\]
increment normalization(\textit{doctor}) by 8.844
increment normalization(\textit{nurse}) by 8.844

<table>
<thead>
<tr>
<th></th>
<th>#1</th>
<th>#2</th>
<th>normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>8.844</td>
<td>0</td>
<td>8.844</td>
</tr>
<tr>
<td>nurse</td>
<td>8.844</td>
<td>0</td>
<td>8.844</td>
</tr>
<tr>
<td>actor</td>
<td>0</td>
<td>######</td>
<td>0</td>
</tr>
</tbody>
</table>

Support and normalization matrix
• Taking doctor and actor

\[ w_1 = \text{doctor} \quad w_3 = \text{actor} \]

\[ c_{1,3} = \text{person} \]

\[ v_{1,3} = 2.005 \]

\[ \text{support (medical doctor)} = \text{support (medical doctor)} + 2.005 \]

\[ \text{support (phd)} = \text{support (phd)} + 2.005 \]

\[ \text{support (performer)} = \text{support (performer)} + 2.005 \]

\[ \text{increment normalization (doctor) by 2.005} \]

\[ \text{increment normalization (actor) by 2.005} \]
<table>
<thead>
<tr>
<th></th>
<th>#1</th>
<th>#2</th>
<th>normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>8.844+2.005</td>
<td>2.005</td>
<td>8.844+2.005</td>
</tr>
<tr>
<td>nurse</td>
<td>8.844</td>
<td>0</td>
<td>8.844</td>
</tr>
<tr>
<td>actor</td>
<td>2.005</td>
<td>#######</td>
<td>2.005</td>
</tr>
</tbody>
</table>
• Taking nurse and actor

\[ w_2 = \text{nurse} \quad w_3 = \text{actor} \]
\[ c_{1,3} = \text{person} \]
\[ v_{2,3} = 2.005 \]

\[
\text{support (health professional)} = \text{support (health professional)} + 2.005
\]
\[
\text{support (nanny)} = \text{support (nanny)} + 2.005
\]
\[
\text{support (performer)} = \text{support (performer)} + 2.005
\]

increment normalization(nurse) by 2.005
increment normalization(actor) by 2.005
<table>
<thead>
<tr>
<th></th>
<th>#1</th>
<th>#2</th>
<th>normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>8.844+2.005</td>
<td>2.005</td>
<td>8.844+2.005</td>
</tr>
<tr>
<td>nurse</td>
<td>8.844+2.005</td>
<td>2.005</td>
<td>8.844+2.005</td>
</tr>
<tr>
<td>actor</td>
<td>2.005+2.005</td>
<td>######</td>
<td>2.005+2.005</td>
</tr>
</tbody>
</table>
Membership function

$$\varphi_{i, k} = \frac{\text{support}[i, k]}{\text{normalization}[i]}$$

$$\varphi_{\text{doctor, medical doctor}} = \frac{\text{support}[\text{doctor, medical doctor}]}{\text{normalization}[\text{doctor}]} = \frac{2.005 + 8.844}{2.005 + 8.844} = 1$$

$$\varphi_{\text{doctor, phd}} = \frac{\text{support}[\text{doctor, phd}]}{\text{normalization}[\text{doctor}]} = \frac{2.005}{2.005 + 8.844} = 0.185$$
Find the most relevant group

$W'$ is the set of words that an ideal human being would cluster together.

$$W' \subseteq \bigcup_{i=1}^{n} S_i$$

Using the above algorithm we can cluster doctor (medical doctor), nurse (health professional), actor (performer) in a cluster.
• On average, Resnik's algorithm achieved approximately 89% of the performance of human annotators performing the same task.
Linking to WordNet using a Bilingual Dictionary

- Multilingual resources are difficult to obtain.
- Associate Chinese vocabulary items with nodes in WordNet using definitions in the CETA Chinese-English dictionary.
- For example, consider the following dictionary entries:
  a) 阿伯
     1. <lit> brother-in-law (husband's elder brother)
     2. <reg> father
     3. <reg> uncle (father's elder brother)
     4. uncle (form of address to an older man)
- 唱旦的 : actress, player of female roles.
• While associating Chinese associations as these with WordNet taxonomy, we must avoid inappropriate senses.

For ex:- 阿伯

in (a) should not be associated with father as in sense of Church's father or priest.

• English glosses in the CETA dictionary are generally too short to take advantage of word overlap in this fashion although they are very similar in meaning but for few cases.

• c): 字盘 1. case i.e., upper case or lower case 2. dial of a watch.

• inspection of the dictionary confirms that when multiple definitions are present they tend more toward polysemy than homonymy.
• Resnik conducted experiments to access the extent to which word sense disambiguation algorithm can be used to identify relevant noun senses in wordnet for chinese words in the CETA dictionary using the english definitions as the source of similar nouns to disambiguate.

• Using some simple heuristics like:-

• Find head nouns i.e. Nouns heading definitons

• for ex: a ) uncle , brother-in-law , father

    b ) actress , player

• Wordnet's noun database was used to automatically find compound nominals .

    For ex: - record player would have the compound record_player as noun rather than player.
Experiment

- Two independent judges were recruited for assistance in annotating the test set: one a native Chinese speaker and the other a Chinese language expert for the US government.
- Judges independently annotated the 100 test items.
- For each item, the judge was given:
  - Chinese word
  - English definition
  - A list of all word net sense description for any head noun in the associated noun group.
Procedure

- For each item the judges were asked if the word was known. If the response was negative, the judge moved to next word.

- For known objects, the instructions were as follows: -
  for each WordNet definition, you will see 6 boxes:

  ![Image of 6 boxes]

- For each definition:
  - if you think the Chinese word can have that meaning, select the number corresponding to your confidence in that choice, where 1 is lowest confidence and 5 is highest confidence.
• If the Chinese word cannot have that meaning, but can have a more specific meaning, select **is-a**.

• For example, if the Chinese word means truck

  "and the WordNet definition is automotive vehicle: self-propelled wheeled vehicle",

  you would select this option. That is, it makes sense to say that this Chinese word describes a concept that is a kind of automotive vehicle. Then pick 1, 2, 3, 4, or 5 as your confidence in this decision.

• If neither of the above cases apply for this WordNet definition, don't check on anything for this definition.
• For ex: -

急讯 - Urgent message or urgent dispatch

• The list contains the following wordnet sense descriptions, as generated by the head nouns message and dispatch. Tick on the appropriate ratings.
  • message, content, subject matter, substance: what a communication that is about something is about.
    
    | 1 | 2 | 3 | 4 | 5 |
    |---|---|---|---|---|
    | s_a |
  
  • dispatch, expedition, expeditiousness, fastness: subconcept of celerity, quick-ness, rapidity.
    
    | 1 | 2 | 3 | 4 | 5 |
    |---|---|---|---|---|
    | s_a |
  
  • dispatch, despatch, communique: an official report usually sent in haste.
    
    | 1 | 2 | 3 | 4 | 5 |
    |---|---|---|---|---|
    | s_a |
• message: a communication usually brief that is written or spoken or signaled; he sent a three-word message
• dispatch, despatch, shipment: the act of sending off something
• dispatch, despatch: the murder or execution of someone
• IS_A was not used in the analysis of results but it provided the judges to give their decision on words that don't have equivalent translation in English.

• For ex:-

  春节 - the spring festival, lunar new year, Chinese new year

• A judge would classify the word as

  festival: a day or period of time set aside for feasting and celebration.

• Since chinese new year does not appear as wordnet concept
• After the test set has been annotated, evaluation is done according to two conditions:

  • Selection – The annotator correctly identified which wordnet senses are to be considered correct.
  • Filtering - The annotator correctly identified which wordnet senses are to be considered incorrect.
• An algorithm being tested against this set must include the senses that were labelled correctly by human annotator.

For ex:-

if the “dispatch, despatch: the murder or execution of someone“ is included by the algorithm, it should be penalized.
• For the selection paradigm, the goal is to identify WordNet senses to include.

• The precision and recall is as follows:

\[ P_{\text{selection}} = \frac{\text{no of correctly included senses}}{\text{no of included senses}} \]

\[ R_{\text{selection}} = \frac{\text{no of correctly included senses}}{\text{no of correct senses}} \]
• For the filtering paradigm, the goal is to identify WordNet senses to exclude.

• The precision and recall is as follows:-

\[
P_{\text{filtering}} = \frac{\text{no of correctly excluded senses}}{\text{no of excluded senses}}
\]

\[
R_{\text{filtering}} = \frac{\text{no of correctly excluded senses}}{\text{no of senses labelled incorrect}}
\]
• In the filtering paradigm, precision begins with the set of senses that the method filtered out and computes the proportion that were correctly filtered out.

• Recall in filtering begins with the set of senses that should have been excluded i.e. the incorrect ones and computes the proportion of these that the method actually managed to exclude.
Results

<table>
<thead>
<tr>
<th></th>
<th>( P_{\text{selection}} )</th>
<th>( R_{\text{selection}} )</th>
<th>( P_{\text{selection}} )</th>
<th>( R_{\text{selection}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>29.5</td>
<td>31.2</td>
<td>88.0</td>
<td>87.1</td>
</tr>
<tr>
<td>Algorithm</td>
<td>36.9</td>
<td>69.9</td>
<td>93.8</td>
<td>79.3</td>
</tr>
<tr>
<td>Judge 2</td>
<td>54.8</td>
<td>55.6</td>
<td>91.9</td>
<td>91.7</td>
</tr>
</tbody>
</table>

Evaluation using Judge 1 as reference standard considering items
Selected with confidence 3 and above.

<table>
<thead>
<tr>
<th></th>
<th>Judge 2</th>
<th>Algorithm</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Include</td>
<td>Exclude</td>
<td>Include</td>
</tr>
<tr>
<td>Judge 1</td>
<td>Include</td>
<td>40</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Exclude</td>
<td>33</td>
<td>363</td>
</tr>
</tbody>
</table>

Agreement and disagreement with judge 1
Conclusions

• Presented semantic similarity measures in an IS-A taxonomy.
• Information Content Measure proved useful and gives good results, very close to human method.
• Semantic similarity measure in case of disambiguating nominal compounds is a good measure.
• In case of Word Sense Selection, the result clearly shows that the algorithm is better than the baseline, but also indicate that it is over-generating senses, which hurts selection precision.
Conclusions

• Below baseline performance in filtering recall suggests the algorithm is choosing to allow senses that it should be filtering out.

• This pattern of results suggests that the best use of this algorithm at its present level of performance would be as a filter for a lexical acquisition process with a human in the loop.
References


Kullback Leibler distance

- It measures the expected number of extra bits required to code samples from $P$ when using a code based on $Q$, rather than using a code based on $P$.
- $KL(P,Q) = \sum_i P(i) \log \left( \frac{P(i)}{Q(i)} \right)$