Word Sense Induction using the SnS method

Phd thesis

Marek Kozłowski

Poznan, 01.04.2014
Contents

1) Ambiguity
2) WSD vs WSI
3) WSI Applications
4) WSI related work
5) The SnS description
6) Exploration of senses by example
7) Qualitative evaluation using SemEval2013 task11 [1]
8) Conclusions
Preface

The times of exponential growth of data
- Mainly unstructured data
- Unlikely to be analyzed by humans
- Currently used approaches are based on lexico-syntactic analysis of text → words occurrences

Two main flaws of the currently used approach are:
- Inability to identify documents using different wordings
- Lack of context-awareness what leads to retrieval documents which are not pertinent to the user needs

Knowledge of an actual meaning of a polysemous word can significantly improve retrieving more relevant documents or extracting relevant information from texts.
Ambiguity

Ambiguity around us
• More than 73% of words in English are polysemous [2]
• Average number of meanings per word is approximately 6
• Apple can be used as a software company name, personal computers, fruit, or river.
• Plant can be used to mean a botanical life form or a industrial building.
• Bass may refer respectively to low-frequency tones and a type of fish.

Sample:
Suppose Robin and Joe are talking, and Joe states, “The bank on the left is solid, but the one on the right is crumbling.” What are Robin and Joe talking about? Are they on Wall Street looking at the offices of two financial institutions, or are they floating down the Mississippi River looking for a place to land their canoe?
Foundations

Word sense theories

- There are many approaches to define sense (referential theory, mentalist, behaviourist and use theory)
- „...the meaning of a word is its use in the language“ (Wittgenstein)
- Distributional hypothesis introduced by Zellig Harris, which can be summarized in a few words: „a word is characterized by the company it keeps“
- **Main assumption used by context-based wsd methods:** Semantically similar words tend to occur in similar contexts.

Sample:
wine: beer, white wine, red wine, Chardonnay, champagne, fruit, food, coffee, juice, Cabernet, cognac, vinegar, Pinot noir, milk, vodka,…
Basic concepts

Word Sense Discovery – ambiguous concept

- **Word sense disambiguation** -
  - process of meaning identification for words in context
  - is the ability to identify the meaning of words in context in a computational manner
  - is an AI-complete problem, that is a problem whose difficulty is equivalent to solving central problems of artificial intelligence.

- **Word sense induction** -
  - subtask of unsupervised WSD
  - task of automatically identifying senses of words in texts, without the need for handcrafted resources or manually annotated data.
Applications

Possible applications of WSI

- **Information retrieval** e.g: new type of Web Search Engines Or Local Domain Oriented Searchers
- **Information Extraction** e.g: acronym expansion, people disambiguation
- **Question Answering** - the main strategy for question answering is to find documents that have the right content even if the same words are not used.
- **Machine Translation** - context to decide about sense of translated word is necessary
- **Lexicographers supporting tool**
- **Support tools for building ontologies** - create relations between concepts and theirs wordings representatives (lexemes)
The WSI approaches can be divided into the following groups

- Clustering context vectors approach ([5], [6]) consists in grouping of the contexts, so that a given target word occurs
- Extended clustering techniques like LSA[7], CBC[8]
- Bayesian methods [9] model the contexts of the ambiguous word as samples from a multinomial distribution over senses (characterized as distributions over words).
- The WSI graph-based approaches ([10], [11], [12]) represent each word co-occurring with the target word (within a context) as a vertex. Two vertices are connected via an edge if they co-occur in one or more contexts of the target word.
- Frequent termsets based algorithms exploit classical data mining methods (as association rule mining) to induce senses ([14], [15]).
Methodology

The main assumptions:
- the granularity problem
- the contextual patterns representation

Concepts

1) Context is

\[ C(P, t) = \{ x | x, t \in \mathcal{O} \land x \neq t \land P \in \mathcal{P} \land x \in P \land \]

\[ (x \in PN \lor ne(x) \lor pos(x) = \text{noun}) \} \]

\[ C(P_1, apple) = \{ \text{album, july 19, american rock, studio album, rock band, record, american, mother love bone, rock music} \} \]
Methodology

2) Contextual pattern are

\[ \mathcal{CP}(\mathcal{C}(t)) = \{X \mid X \in F(t) \text{ and } \exists Y \supset X \text{ with } \text{supp}(X) = \text{sup}(Y)\} \]

\( \mathcal{CP}_1: \{\text{album, release, bass guitar, single}\} \text{ with } \text{sup} = 2 \)

3) Sense Frame are

\[ \mathcal{SF}(t) = \text{stree}(\emptyset, t) = \langle (mp_1, \text{stree}(mp_1, t)), ..., (mp_n, \text{stree}(mp_n, t)) \rangle = \langle sf_{1,t}, ..., sf_{n,t} \rangle \text{ where} \]

\[ (mp_i, \text{stree}(mp_i, t)) \text{ is denoted as } sf_{i,t} \text{ and } i=1,..,n \]

4) Sense is a set of clustered sense frames
Algorithm

Main Algorithm – steps done to find senses for a query
Input: set of documents (wikipedia, or another corpora), which is indexed by Lucene
- Using full-text search there are found all paragraphs of documents which contains a given query
- Simple context build stage → only alphanumeric noun-phrases and proper names surrounding the given term are persisted as a contexts
- Relevant patterns build stage → closed frequent sets mining on contexts is taken to build set of contextual patterns
- Sense frames are constructed from contextual patterns (subset-superset relations)
- Sense are generated from clustered sense frames

Output for end-user is a set of senses, with multi-level hierarchies and matching them documents.
The SnS method

- The SenseSearcher (SnS) is a word sense induction algorithm based on closed frequent sets and multi-level sense representation. SnS performed better than methods using vector space modelling. Induced senses by SnS characterize better readability (are more intuitive), also they are hierarchical, what gives them flexible granularity.

- Key features of SnS are:
  - ability to find infrequent, dominated senses;
  - number of likely senses determined by content of corpora, there is no fixed threshold determining constant number of retrieved senses;
  - multi-level hierarchies of senses (describing subsenses)
Evaluation

First, we evaluate sense exploration in a large textual corpus. We will show in the experiments that SnS is enable to build hierarchical sense representations and identify dominated, infrequent senses. SnS was evaluated using Wikipedia corpus.

Second, we test SnS as a web search result clustering WSI-based algorithm. These experiments aimed at comparing SnS with other WSI algorithms within 2013 SemEval Task no 11 "Word Sense Induction within an End-User Application". According to the SemEval standards we evaluate SnS with others by means of a broad scope of quality/diversification measures.
Sample evaluation

<table>
<thead>
<tr>
<th>Context</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer</td>
<td>99</td>
</tr>
<tr>
<td>music</td>
<td>35</td>
</tr>
<tr>
<td>county</td>
<td>31</td>
</tr>
<tr>
<td>others</td>
<td>0</td>
</tr>
<tr>
<td>state</td>
<td>19</td>
</tr>
<tr>
<td>population</td>
<td>10</td>
</tr>
<tr>
<td>others</td>
<td>0</td>
</tr>
<tr>
<td>census</td>
<td>8</td>
</tr>
<tr>
<td>area</td>
<td>8</td>
</tr>
<tr>
<td>community</td>
<td>8</td>
</tr>
<tr>
<td>valley</td>
<td>5</td>
</tr>
<tr>
<td>festival</td>
<td>5</td>
</tr>
<tr>
<td>fruit</td>
<td>30</td>
</tr>
<tr>
<td>cultivar</td>
<td>12</td>
</tr>
<tr>
<td>is the</td>
<td>10</td>
</tr>
<tr>
<td>new</td>
<td>5</td>
</tr>
<tr>
<td>golden</td>
<td>18</td>
</tr>
<tr>
<td>film</td>
<td>17</td>
</tr>
<tr>
<td>book</td>
<td>13</td>
</tr>
</tbody>
</table>

131) Apple of Discord. An apple (symbolism) apple of discord is a reference to the Golden Apple of Discord () which, according to Greek mythology, the goddess Eris (mythology) Eris (Greek language Gr. ἔρις, "Strife") said that she would give "to the fairest" at the wedding of Peleus and Thetis, sparking a vanity-fueled dispute between Hera, Athena and Aphrodite that eventually led to the Trojan War (For the complete story, see The Judgement of Paris (mythology) Judgement of Paris). Thus, "apple of discord" became a euphemism for the core, kernel, or crux of an argument, or for a small matter that could lead to a bigger dispute. The Ancient Greek word "μῆλον" also means sheep or goat. Homer in his Odyssey describes how Odysseus sacrificed two goats (μῆ..."

184) Cameo (apple). The Cameo is a cultivar of apple, discovered by chance by the Caudle Family in a Dryden, Washington Dryden, Washington orchard in 1987. Its parentage is uncertain; it may be a cross between a Red Delicious and a Golden Delicious.

Figure C.1: Senses for **apple** extracted from Wikipedia articles
Sample evaluation

Figure C.2: Senses for bass extracted from Wikipedia articles
### Context

<table>
<thead>
<tr>
<th>Context</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>specie</td>
<td>317</td>
</tr>
<tr>
<td>cartoon</td>
<td>29</td>
</tr>
<tr>
<td>others</td>
<td>0</td>
</tr>
<tr>
<td>jerry</td>
<td>18</td>
</tr>
<tr>
<td>film</td>
<td>17</td>
</tr>
<tr>
<td>short</td>
<td>15</td>
</tr>
<tr>
<td>animated</td>
<td>13</td>
</tr>
<tr>
<td>character</td>
<td>9</td>
</tr>
<tr>
<td>computer</td>
<td>15</td>
</tr>
<tr>
<td>others</td>
<td>0</td>
</tr>
<tr>
<td>is the</td>
<td>5</td>
</tr>
<tr>
<td>user</td>
<td>5</td>
</tr>
<tr>
<td>computing</td>
<td>4</td>
</tr>
<tr>
<td>time</td>
<td>4</td>
</tr>
<tr>
<td>child</td>
<td>13</td>
</tr>
<tr>
<td>others</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>4</td>
</tr>
<tr>
<td>band</td>
<td>9</td>
</tr>
<tr>
<td>others</td>
<td>0</td>
</tr>
<tr>
<td>rock</td>
<td>6</td>
</tr>
</tbody>
</table>

1) Mouse. A mouse (plural: mice) is a small mammal belonging to the order of rodents. The best known mouse species is the common house mouse (Mus musculus). It is also a popular pet. In some places, certain kinds of Apodemus field mice are also common. This rodent is eaten by large birds such as hawks and eagles. They are known to invade homes for food and occasionally shelter.

8) Algerian mouse. The Algerian mouse, or western Mediterranean mouse, (Mus spretus) is a wild species of mouse closely related to the house mouse, native to open habitats around the weste...

11) Country Mouse. The Country Mouse (Pseudomys patrius) also known as the Pebble-Mound Mouse or Eastern Pebble Mound Mouse is a species of rodent in the Family Muridae. It is found only in Australia. It is considered to be a rare mouse and was first discovered by Thomas and Dollman i...

---

**Figure C.4:** Senses for *mouse* extracted from Wikipedia articles
Comparison with others

Some of algorithms which were experimentally used as alternatives:

- **LSA – Latent Semantic Analysis**
  - Co-occurring terms are mapped to the same dimensions, not co-occurring terms are mapped to different dimensions
  - Lower number of dimensions leads to generalizations over the simple frequency data.

- **LDA – Latent Dirichlet Allocation**
  - Probabilistic model of text generation
  - Models each document using a mixture over K topics, which are in turn characterized as distributions over words

- **SenseClusters**
  - Perl programs that allows a user to cluster similar contexts together using unsupervised knowledge-lean methods.
  - Second-order context representation
Comparison with LSI

- LSI models the meaning of words and documents by projecting them into a vector space of reduced dimensionality by applying singular value decomposition

1. \[ 0.263\text{“computer”} + 0.257\text{“inc”} + 0.201\text{“macintosh”} + 0.159\text{“system”} + 0.152\text{“software”} + 0.145\text{“product”} + 0.142\text{“model”} + 0.130\text{“use”} + 0.124\text{“mac”} + 0.120\text{“line”} \]

2. \[ 0.367\text{“county”} + 0.325\text{“state”} + 0.294\text{“population”} + 0.280\text{“census”} + 0.201\text{“river”} + 0.175\text{“davies”} + 0.169\text{“valley”} + 0.151\text{“community”} + 0.139\text{“town”} + 0.131\text{“place”} \]

3. \[ 0.215\text{“system”} + -0.195\text{“series”} + 0.188\text{“valley”} + -0.166\text{“model”} + -0.157\text{“product”} + 0.151\text{“file”} + 0.148\text{“program”} + 0.142\text{“mac”} + 0.139\text{“district”} + 0.135\text{“operating”} \]

4. \[ -0.296\text{“album”} + -0.235\text{“band”} + -0.212\text{“rock”} + -0.201\text{“record”} + 0.181\text{“music”} + 0.140\text{“population”} + 0.137\text{“census”} + -0.134\text{“pie”} + 0.130\text{“county”} + -0.118\text{“tree”} \]

5. \[ -0.218\text{“album”} + 0.201\text{“cultivar”} + -0.198\text{“pathogenic”} + -0.171\text{“band”} + -0.167\text{“plant”} + -0.161\text{“rock”} + -0.155\text{“virus”} + -0.149\text{“flexiviridae”} + -0.146\text{“family”} + 0.132\text{“fruit”} \]
WSI within End-user Application

Evaluation of WSI methods is difficult because there is no easy way to compare and rank various representations of senses. Evaluating WSI methods is actually a special case of a more general and difficult problem of evaluating clustering algorithms. In order to find out more rigid ways to compare results of sense induction systems, Navigli and Vannella [1] organized Semeval-2013 task 11.

The task is stated as follows:

given a target query, induction and disambiguation systems are requested to cluster and diversify the search results returned by a search engine for that query
In order to perform comparisons with SemEval systems we customized SnS by adding the results clustering and diversication phase. The clustering is performed in two phases: (1) simultaneously during sense induction, and (2) after sense discovering clustering the results that remained not grouped before.

*Each sense frame has the main contextual pattern, so according to sense frames the snippets containing the main pattern are grouped in the corresponding result cluster. Non-grouped snippets are tested iteratively against each of the induced sense. The similarity measure is defined as intersection cardinality between the snippet and sense cluster's bag-of-words. Within each cluster the snippets are sorted using this measure. Clusters are sorted by the support of their sense frames.*
Compared systems

SemEval task 11 participating systems

- HDP systems adopt WSI methodology based on a non-parametric model using Hierarchical Dirichlet Process. Systems are trained over extracts from the full text of English Wikipedia.
- Satty-approach system implements the idea of monotone submodular function optimization, using a greedy algorithm.
- UKP systems exploit graph-based WSI methods and external resources (Wikipedia or ukWaC).
- Duluth systems are based on second-order context clustering as provided in SenseClusters, a freely available open source software package.
- Rakesh system exploits external sense inventories for performing the disambiguation task. It employs YAGO hierarchy and DBPedia in order to assign senses to the search results.
Scoring

Clustering quality measures

- Rand Index (RI)

\[ RI(C, G) = \frac{TP + TN}{TP + FP + FN + TN} \]

- Adjusted Rand Index (ARI)

\[ ARI(C, G) = \frac{RI(C, G) - E(RI(C, G))}{\max RI(C, G) - E(RI(C, G))} \]

- Jaccard Index (JI)

\[ JI(C, G) = \frac{TP}{TP + FP + FN} \]

- Precision

\[ P(C_j) = \frac{|C_j^t|}{|C_j|} \]

- Recall

\[ R(t) = \frac{|\bigcup_{C_j \in c^t} C_j^t|}{n_t} \]

- F-measure

\[ P = \frac{\sum_{C_j \in c} P(C_j)|C_j|}{\sum_{C_j \in c} |C_j|}; \quad R = \frac{\sum_{t \in T} R(t)n_t}{\sum_{t \in T} n_t} \]

Diversification quality measures

- S-recall@K

\[ S\text{-recall}@K = \frac{|\bigcup_{i=1}^{K} \text{subtopics}(r_i)|}{m} \]

- S-precision@r

\[ S\text{-precision}@r = \frac{|\bigcup_{i=1}^{K_r} \text{subtopics}(r_i)|}{K_r} \]
## Results – quality measures

**Table 1.** The results of clustering experiments on SEMEVAL data set (in %).

<table>
<thead>
<tr>
<th>Type</th>
<th>System</th>
<th>RI</th>
<th>ARI</th>
<th>JI</th>
<th>F1</th>
<th>cl.</th>
<th>ACS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSI</td>
<td>HDP-CLS-LEMMA</td>
<td>65.22</td>
<td>21.31</td>
<td>33.02</td>
<td><strong>68.30</strong></td>
<td>6.63</td>
<td>11.07</td>
</tr>
<tr>
<td></td>
<td>HDP-CLS-NOLEMMA</td>
<td>64.86</td>
<td><strong>21.49</strong></td>
<td>33.75</td>
<td>68.03</td>
<td>6.54</td>
<td>11.68</td>
</tr>
<tr>
<td></td>
<td>SATTY-APPROACH1</td>
<td>59.55</td>
<td>7.19</td>
<td>15.05</td>
<td>67.09</td>
<td>9.90</td>
<td>6.46</td>
</tr>
<tr>
<td></td>
<td>DULUTH SYS9.PK2</td>
<td>54.63</td>
<td>2.59</td>
<td>22.24</td>
<td>57.02</td>
<td>3.32</td>
<td>19.84</td>
</tr>
<tr>
<td></td>
<td>DULUTH SYS1.PK2</td>
<td>52.18</td>
<td>5.74</td>
<td>31.79</td>
<td>56.83</td>
<td>2.53</td>
<td>26.45</td>
</tr>
<tr>
<td></td>
<td>DULUTH SYS7.PK2</td>
<td>52.04</td>
<td>6.78</td>
<td>31.03</td>
<td>58.78</td>
<td>3.01</td>
<td>25.15</td>
</tr>
<tr>
<td></td>
<td>UKP-WSI-WP-LLR2</td>
<td>51.09</td>
<td>3.77</td>
<td>31.77</td>
<td>58.64</td>
<td>4.17</td>
<td>21.87</td>
</tr>
<tr>
<td></td>
<td>UKP-WSI-WP-PMI</td>
<td>50.50</td>
<td>3.64</td>
<td>29.32</td>
<td>60.48</td>
<td>5.86</td>
<td>30.30</td>
</tr>
<tr>
<td></td>
<td>UKP-WSI-WACKY-LLR</td>
<td>50.02</td>
<td>2.53</td>
<td><strong>33.94</strong></td>
<td>58.26</td>
<td>3.64</td>
<td>32.34</td>
</tr>
<tr>
<td>WSD</td>
<td>RAKESH</td>
<td><strong>58.76</strong></td>
<td>8.11</td>
<td><strong>30.52</strong></td>
<td><strong>39.49</strong></td>
<td>9.07</td>
<td>2.94</td>
</tr>
<tr>
<td>SNS</td>
<td>SNS</td>
<td><strong>65.84</strong></td>
<td><strong>22.19</strong></td>
<td><strong>34.26</strong></td>
<td><strong>70.16</strong></td>
<td>8.82</td>
<td>8.46</td>
</tr>
</tbody>
</table>
## Results – diversification measures

### Table 2. The results for S-recall@K and S-precision@r on SEMEVAL data set (in %).

<table>
<thead>
<tr>
<th>Type</th>
<th>System</th>
<th>S-recall</th>
<th></th>
<th></th>
<th></th>
<th>S-precision</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>K=5</td>
<td>K=10</td>
<td>K=20</td>
<td>K=40</td>
<td>r=50</td>
<td>r=60</td>
<td>r=70</td>
<td>r=80</td>
<td></td>
</tr>
<tr>
<td>WSI</td>
<td>HDP-CLS-LEMMA</td>
<td>48.13</td>
<td>65.51</td>
<td>78.86</td>
<td>91.68</td>
<td>48.85</td>
<td>42.93</td>
<td>35.19</td>
<td>27.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HDP-CLS-NOLEMMA</td>
<td>50.80</td>
<td>63.21</td>
<td>79.26</td>
<td>92.48</td>
<td>48.18</td>
<td>43.88</td>
<td>34.85</td>
<td>29.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SATTY-APPROACH1</td>
<td>38.97</td>
<td>48.90</td>
<td>62.72</td>
<td>82.14</td>
<td>34.94</td>
<td>26.88</td>
<td>23.55</td>
<td>20.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DULUTH.SYS9.PK2</td>
<td>37.15</td>
<td>49.90</td>
<td>68.91</td>
<td>83.65</td>
<td>35.90</td>
<td>29.72</td>
<td>25.26</td>
<td>21.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DULUTH.SYS1.PK2</td>
<td>37.11</td>
<td>53.29</td>
<td>71.24</td>
<td>88.48</td>
<td>40.08</td>
<td>31.31</td>
<td>26.73</td>
<td>24.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DULUTH.SYS7.PK2</td>
<td>38.88</td>
<td>53.79</td>
<td>70.38</td>
<td>86.23</td>
<td>39.11</td>
<td>30.42</td>
<td>26.54</td>
<td>23.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UKP-WSI-WP-LLR2</td>
<td>41.07</td>
<td>53.76</td>
<td>68.87</td>
<td>85.87</td>
<td>42.06</td>
<td>32.04</td>
<td>26.57</td>
<td>22.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UKP-WSI-WP-PMI</td>
<td>40.45</td>
<td>56.25</td>
<td>68.70</td>
<td>84.92</td>
<td>42.83</td>
<td>33.40</td>
<td>26.63</td>
<td>22.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UKP-WSI-WACKY-LLR</td>
<td>41.19</td>
<td>55.41</td>
<td>68.61</td>
<td>83.90</td>
<td>42.47</td>
<td>31.73</td>
<td>25.39</td>
<td>22.71</td>
<td></td>
</tr>
<tr>
<td>WSD</td>
<td>RAKESH</td>
<td>46.48</td>
<td>62.36</td>
<td>78.66</td>
<td>90.72</td>
<td>48.00</td>
<td>39.04</td>
<td>32.72</td>
<td>27.92</td>
<td></td>
</tr>
<tr>
<td>SNS</td>
<td>SNS</td>
<td>47.36</td>
<td>62.96</td>
<td>74.10</td>
<td>87.79</td>
<td>47.95</td>
<td>37.99</td>
<td>31.68</td>
<td>24.10</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

**Summarization:** SnS is a novel WSI knowledge-poor algorithm SnS, based on text mining approaches, namely closed frequent termsets. It uses small or medium size text corpus in order to identify senses. It converts simple contexts into relevant contextual patterns. Using patterns SnS build hierarchical structures called sense frames. Discovered sense frames usually are independent senses, but sometimes (e.g. because of too small corpus) can point the same sense. Finally using clustering methods sense frames are grouped in order to find similar ones referring to the same main sense.

**Applications:** semantic search engines, machine translation, tools for lexicographers

**Future work:**.....
References

References