Automated word puzzle generation using topic models and semantic relatedness measures

Balázs Pintér, Gyula Vörös, Zoltán Szabó and András Lőrincz

ELTE IK

2012. 02. 11.
1 Introduction
- Our goal
- The method

2 Steps of the algorithm
- Modeling the corpus as a combination of latent topics
- Identifying consistent sets
- Generating the puzzles

3 Results
- The performance of the three topic models
- Some interesting puzzles
Word puzzles
- Are used in education, psychometry, etc. (e.g., TOEFL)
- Are costly to design and maintain

Our goal is to generate word puzzles from unstructured and unannotated corpora

Puzzle types
- Odd one out: salmon, shark, whale, elephant
- Choose the related word: regiment, battalion, army | infantry, service, king
- Separate the topics: water, heat, temperature, pressure | superman, clark, luthor, kryptonite
Building blocks of puzzles
- *Consistent sets* (sets of related words): \{salmon, shark, whale\}
- Less related words: elephant

Steps of the algorithm
- Model the corpus as a combination of latent topics
- Identify consistent sets from among these topics
- Generate the puzzles by mixing these sets with less related elements

Building blocks of the algorithm
- Topic models
- Semantic similarity measures
- Network flow
The government introduced a new voice controlled system that allows air force pilots to control aircraft by issuing commands in natural language...
Topic models used

**Latent Semantic Analysis**

\[
\arg \min_{\text{rank}(\hat{X}) = d} \|X - \hat{X}\|_F = USV^T. \tag{1}
\]

**Online Group-Structured Dictionary Learning**

\[
\min_{D, \{\alpha_i\}^M_{i=1}} \sum_{j=1}^M \left( \frac{i}{M} \right)^\rho \sum_{i=1}^M \left( \frac{i}{M} \right)^\rho \left[ \frac{1}{2} \|x_i - D\alpha_i\|^2_2 + \kappa \Omega(\alpha_i) \right] (\kappa > 0), \tag{2}
\]

\[
\Omega(\alpha) = \left( \sum_j \|\alpha_G|_2 \right)^{1/\eta}, \tag{3}
\]

**Latent Dirichlet Allocation**

\[
P(W, Z, \theta, \phi|\alpha, \beta) = \prod_{i=1}^K P(\phi_i|\beta) \prod_{j=1}^M P(\theta_j|\alpha) \prod_{t=1}^{N_j} P(z_j,t|\theta_j)P(w_j,t|\phi_{z_j,t}), \tag{4}
\]
**Topic models used**

**Latent Semantic Analysis**

\[
\arg \min_{\text{rank}(\hat{X})=d} \|X - \hat{X}\|_F = \hat{U} \hat{S} \hat{V}^T.
\] (1)

**Online Group-Structured Dictionary Learning**

\[
\min_{D_i, \{\alpha_i\}_{i=1}^M} \frac{1}{\sum_{j=1}^M (j/M)^\rho} \sum_{i=1}^M \left( \frac{i}{M} \right)^\rho \left[ \frac{1}{2} \|x_i - D \alpha_i\|_2^2 + \kappa \Omega(\alpha_i) \right] \quad (\kappa > 0),
\] (2)

\[
\Omega(\alpha) = \left( \sum_j \|\alpha g_j\|_2^\eta \right) \frac{1}{\eta},
\] (3)

**Latent Dirichlet Allocation**

\[
P(W, Z, \theta, \phi | \alpha, \beta) = \prod_{i=1}^K P(\phi_i | \beta) \prod_{j=1}^M P(\theta_j | \alpha) \prod_{t=1}^{N_j} P(z_j, t | \theta_j) P(w_j, t | \phi_{z_j, t}),
\] (4)
Identifying consistent sets

(a) A consistent set

(b) An inconsistent set
Generating the puzzles

- **Odd one out**
  - Mix a consistent set and a less related word
    - *salmon, shark, whale, elephant*

- **Choose the related word**
  - Mix a consistent set with some less related words
  - Present the words in a different grouping
    - *regiment, battalion, army | infantry, service, king*

- **Separate the topics**
  - Mix two (or more) consistent sets
    - *water, heat, temperature, pressure | superman, clark, luthor, kryptonite*
The performance of the three topic models

(c) Wikipedia

(d) NIPS proceedings
### Some interesting odd one out puzzles

<table>
<thead>
<tr>
<th>Consistent set of words</th>
<th>Odd one out</th>
</tr>
</thead>
<tbody>
<tr>
<td>cao, superman, devil, egypt, singh, language</td>
<td>king, batman, body, bishop, delhi, sound, orbit, view, ancient, image, physical</td>
</tr>
<tr>
<td>wei, clark, demon, egyptian, guru, dialect, force, speech</td>
<td></td>
</tr>
<tr>
<td>liu, luthor, hell, alexandria, sikh, linguistic, motion, hearing, pericles, format</td>
<td></td>
</tr>
<tr>
<td>emperor, kryptonite, soul, pharaoh, saini, spoken, velocity, sound, corinth, compression, equation</td>
<td></td>
</tr>
</tbody>
</table>