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.

Nemzeti Fejlesztési Ügynökség
www.ujszechenyiterv.gov.hu
06 40 638 638

MAGYARORSZÁG MEGÚJUL

The Project is supported by the European Union and co-financed by the European Social Fund (grant agreement no. TAMOP 4.2.1./B-09/1/KMR-2010-0003).

Table of contents

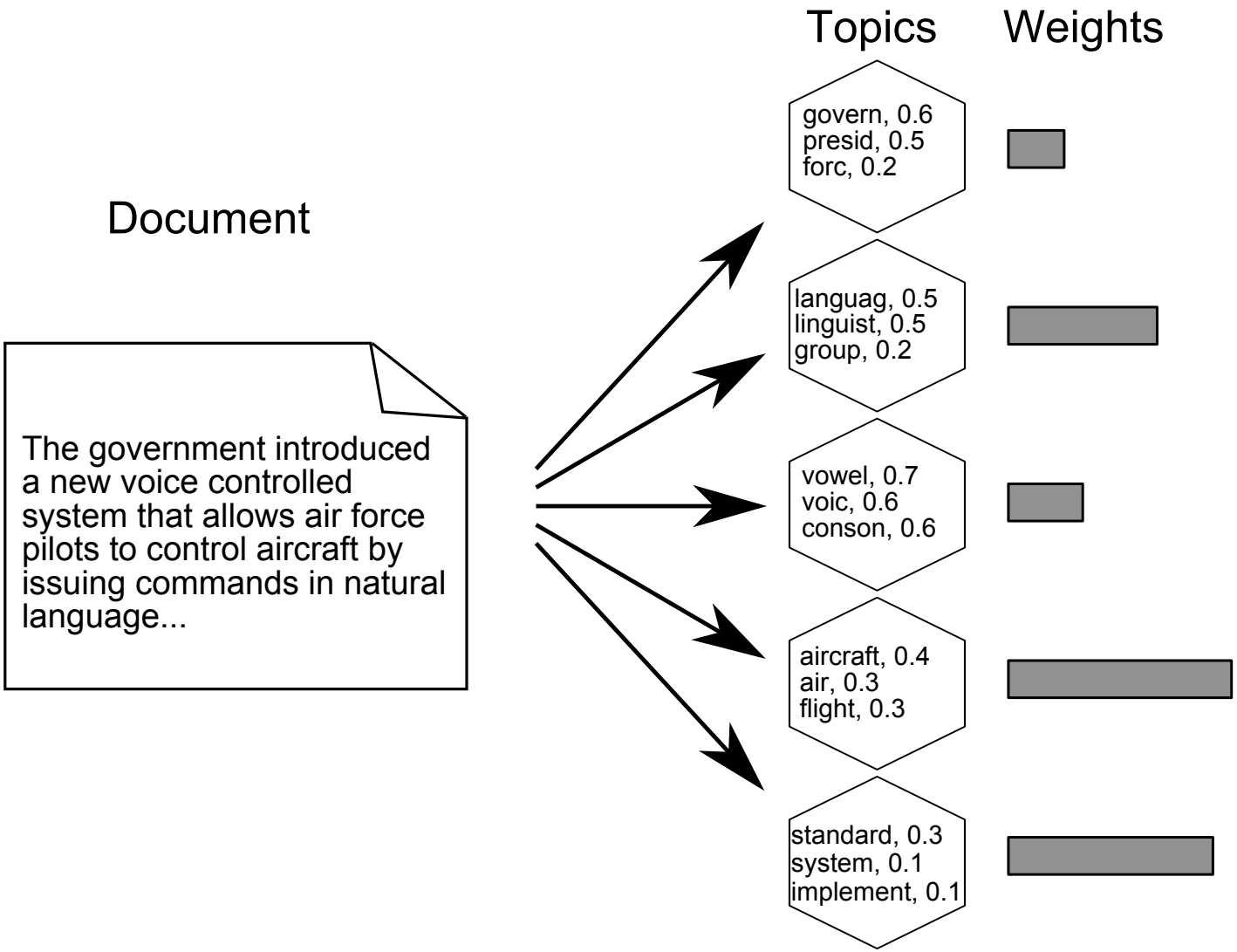
Our goal

- 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*

The method

- 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

Topic models



Topic models used

Latent Semantic Analysis

$$\underset{\mathbf{rank}(\hat{\mathbf{X}})=d}{\arg\min} \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_F = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T. \tag{1}$$

Online Group-Structured Dictionary Learning

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

$$\Omega(\boldsymbol{\alpha}) = \left(\sum_j \left\|\boldsymbol{\alpha}_{\mathbf{G_j}}\right\|_2^{\eta}\right)^{\frac{1}{\eta}}, \tag{3}$$

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

$$\underset{\mathbf{rank}(\hat{\mathbf{X}})=d}{\arg\min} \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_F = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T. \tag{1}$$

Online Group-Structured Dictionary Learning

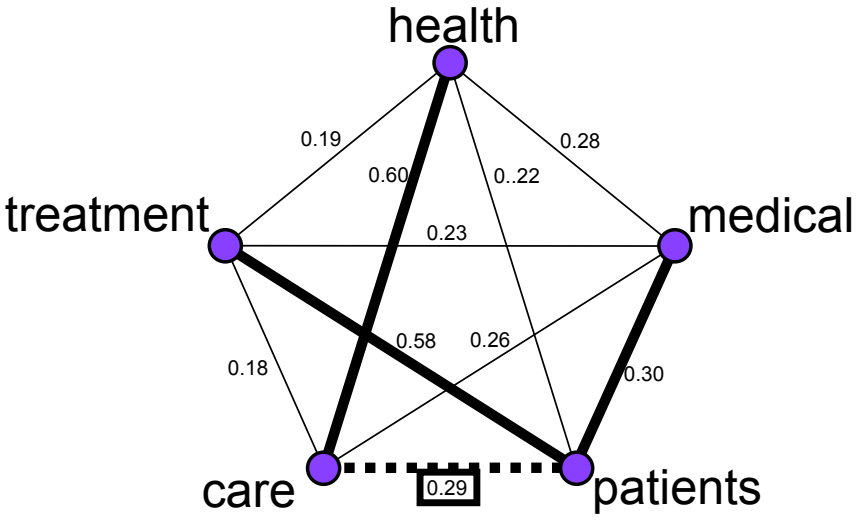
$$\min_{\mathbf{D},\{\boldsymbol{\alpha_i}\}_{i=1}^{M}} \frac{1}{\sum_{j=1}^{M}(j/M)^{\rho}} \sum_{i=1}^{M} \left(\frac{i}{M}\right)^{\rho} \left[\frac{1}{2}\|\mathbf{x_i} - \mathbf{D}\boldsymbol{\alpha_i}\|_2^2 + \kappa\Omega(\boldsymbol{\alpha_i})\right] (\kappa > 0), \tag{2}$$

$$\Omega(\boldsymbol{\alpha}) = \left(\sum_j \left\|\boldsymbol{\alpha}_{\mathbf{G_j}}\right\|_2^{\eta}\right)^{\frac{1}{\eta}}, \tag{3}$$

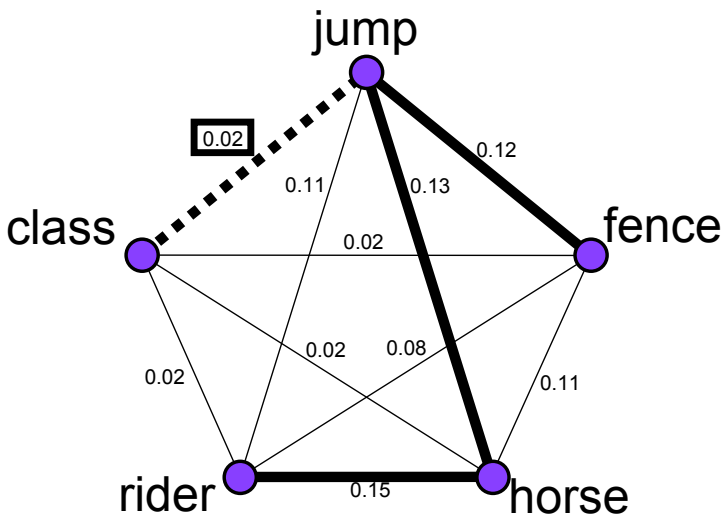
Latent Dirichlet Allocation

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = \prod_{i=1}^{K} P(\phi_i|\beta) \prod_{j=1}^{M} P(\theta_j|\alpha) \prod_{t=1}^{N_j} P(z_{j,t}|\theta_j) P(w_{j,t}|\phi_{z_{j,t}}), \tag{4}$$

Identifying consistent sets



(a) A consistent set

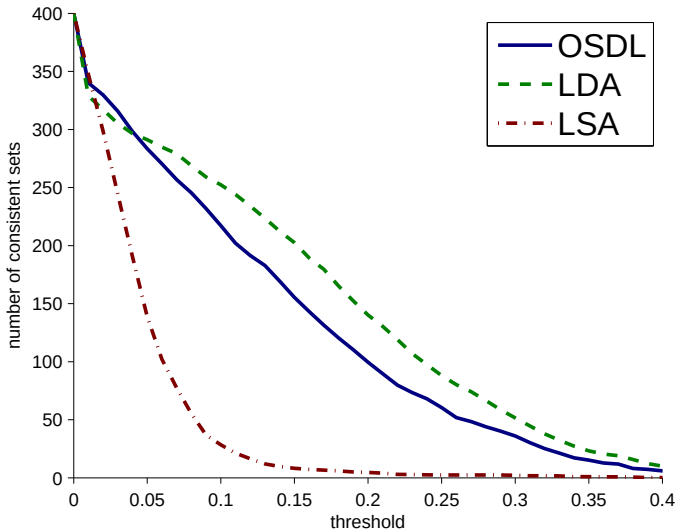


(b) An inconsistent set

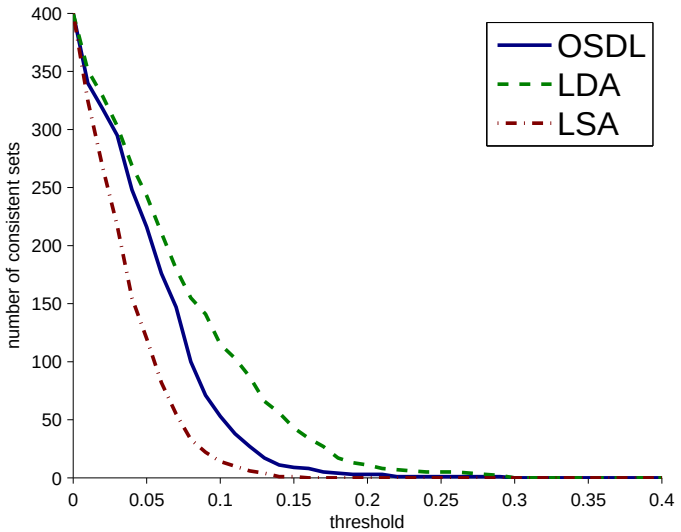
Generating the puzzles

- 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

Consistent set of words				Odd one out
cao	wei	liu	emperor	king
superman	clark	luthor	kryptonite	batman
devil	demon	hell	soul	body
egypt	egyptian	alexandria	pharaoh	bishop
singh	guru	sikh	saini	delhi
language	dialect	linguistic	spoken	sound
mass	force	motion	velocity	orbit
voice	speech	hearing	sound	view
athens	athenian	pericles	corinth	ancient
data	file	format	compression	image
function	problems	polynomial	equation	physical