

# Automated Word Puzzle Generation via Topic Dictionaries

Balázs Pintér, Vörös Gyula, Zoltán Szabó, András Lőrincz

Eötvös Loránd University,  
Budapest, Hungary

ICML – Sparsity, Dictionaries and  
Projections in ML and Signal Processing

June 30, 2012

- Motivation: (word) puzzles – many exciting applications.
- Goal: automated word puzzle generation.
- Method with 3 general components:
  - 1 unlabeled document collection (corpus),
  - 2 topic model → topic dictionary,
  - 3 semantic similarity of word pairs.
- Illustrations.

Many exciting potentials [Verguts and Boeck, 2000]:

- to test/improve skills:
  - IQ test: odd one out puzzle [Carter, 2005],
  - language skills, verbal aptitude,
  - TOEFL: multiple-choice synonym questions.
- in game content generation (computer-, video games).



# Generating Puzzles – Challenges

- Generating/maintaining puzzles: challenging + expensive.
- Central problem:
  - Variety (odd one out puzzles):



# Generating Puzzles – Challenges

- Generating/maintaining puzzles: challenging + expensive.
- Central problem:
  - Variety (odd one out puzzles):



# Generating Puzzles – Challenges

- Generating/maintaining puzzles: challenging + expensive.
- Central problem:
  - Variety (odd one out puzzles):



- Languages are continuously changing (word puzzles):
  - new words are created (e.g., on blogs),
  - existing ones get new meanings ('chat'),
  - words go out of common use ('videotape').
- Automated generation schemes:
  - $\exists$  in quite special cases: sudoku, mazes on chessboards, ...

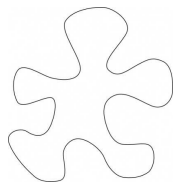
# Generating Puzzles – Challenges-2

- Languages are continuously changing (word puzzles):
  - new words are created (e.g., on blogs),
  - existing ones get new meanings ('chat'),
  - words go out of common use ('videotape').
- Automated generation schemes:
  - $\exists$  in quite special cases: sudoku, mazes on chessboards, ...
  - could be of considerable benefit.





- Automated generation of *word puzzles*: novel field.
- [Colton, 2002]:
  - special puzzles,
  - based on highly structured datasets,
  - requires serious human annotation effort!



- We present an automated word puzzle generation method based on 3 general components [Pintér et al., 2012]:
  - 1 simple document collection (corpus),
  - 2 topic model (LSA, LDA, SDL, ...),
  - 3 semantic similarity of word pairs.
- Capable of generating puzzles of *many different types*:
  - odd one out, choose the related word, separate the topics.
- Can create *domain-specific* puzzles (replace the corpus).
- Parameterizable levels (beginners, ...).

# Puzzle Generator – High-Level View

Three steps:

- 1 Corpus (**X**)  $\xrightarrow{\text{topic model}}$  Topic dictionary: sets of words (**D**).
- 2 Topics (**D**)  $\xrightarrow{\text{discard junk topics}}$  Consistent sets of related words (**C**).
- 3 Consistent Sets (**C**)  $\xrightarrow{\text{add dissimilar items (e.g., words)}}$  Puzzles.

# Step-1 (Corpus $\rightarrow$ Topic Dictionary)

- Corpus:  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M] \in \mathbb{R}^{N \times M}$ .
- Documents:  $\mathbf{x}_m \in \mathbb{R}^N =$  weights assigned to words.
- Topic model ( $\mathcal{T}$ )  $\Rightarrow$  *dictionary*  $\mathbf{D} = \mathcal{T}(\mathbf{X}) = [\mathbf{d}_1, \dots, \mathbf{d}_K] \in \mathbb{R}^{N \times K}$ .

Examples:

- LSA [Deerwester et al., 1990]:  $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$  (SVD)  $\mapsto \mathbf{D} = \mathbf{U}$ .
- SDL (structured sparsity: [Baraniuk et al., 2010, Bach et al., 2012]):

$$\min_{\mathbf{D}} \frac{1}{M} \sum_{i=1}^M \min_{\alpha_i} \left[ \frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \kappa\Omega(\alpha_i) \right]. \quad (1)$$

- LDA [Blei et al., 2003]:  $\mathbf{d}_i$ s with Dirichlet prior  $\mapsto \mathbf{D}$ .

## Step-2 (Topics → Consistent Sets)

- We keep only the  $k$  most significant words of  $\mathbf{d}_i$ s ( $m_i$ ).
- Can still contain junk topics (not related words)  
[Alsumait et al., 2009].
- Goal: determine the consistency of the  $m_i$  sets.
- A word weakly connected to the others: ambiguous puzzle.
- Consistency:  $\Leftarrow$  word least related to the others.

## Step-2 (Topics → consistent sets)

- Semantic relatedness of words ( $w, w'$ ), we use ESA [Gabrilovich and Markovitch, 2009]:

$$s_{ww'} = \cos(\varphi_{ESA}(w), \varphi_{ESA}(w')). \quad (2)$$

=tfidf similarity based on Wikipedia articles.

- Not perfectly accurate:
  - false positives (negatives): high (low) estimated similarity.
  - false negatives: requiring all word pairs to be related  $\neq$  good criteria.

## Step-2 (Topics → Consistent Sets)

DEF: Set  $m$  is consistent, iff

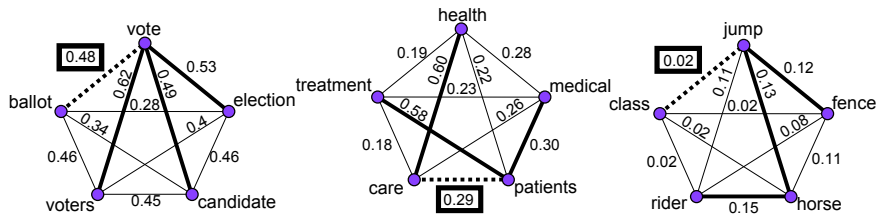
$$\min_{(i,j) \in m \times m, i \neq j} \text{sim}(i, j) := \max_{\text{path}(i,j) \in G := (m, \mathbf{S}|_m)} \min_{e \in \text{path}(i,j)} s_e > \delta. \quad (3)$$

In words:

- Robustness to false negatives: *all paths* ( $\max_{\text{path}(i,j)}$ ),
- Robustness to false positives: *minimum relatedness* ( $\min_e s_e$ ),
- Two most dissimilar words in the set: ( $\min_{i \neq j} \text{sim}(i, j)$ ).

Computation: unique path in the maximal spanning tree of  $G$  [Jungnickel, 2007].

## Step-2 (Topics $\rightarrow$ Consistent Sets) – Example



- Left: highly consistent set ( $\forall \leftrightarrow$  'vote').
- Center: consistent set;  $S_{care,treatment} <$  expected, the method is robust.
- Right: inconsistent set; 'class'  $\nleftrightarrow$  others.



## Step-3 (Consistent Sets $\rightarrow$ Puzzles)

Example (Odd one out puzzle):

**Input:** consistent sets  $\mathcal{C}$ , minimal (maximal) relatedness to consistent sets  $\eta_1$  ( $\eta_2$ )

**for all**  $C \in \mathcal{C}$  **do**

**repeat**

    select random word  $w$

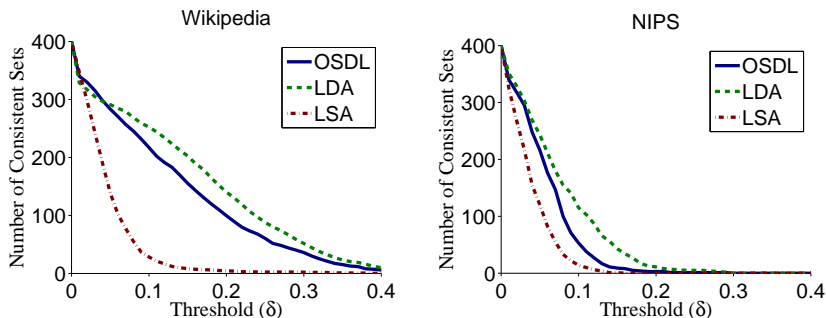
$\sigma \leftarrow \max_{t \in C} s_{wt}$  % max. relatedness of  $w$  to  $C$

**until**  $\eta_1 < \sigma < \eta_2$

  output ( $C, w$ ) puzzle

# Illustration: Number of Consistent Sets

- Methods ( $\mathcal{T}$ ): LSA, LDA, OSDL [Szabó et al., 2011].
- Corpora ( $\mathbf{X}$ ): Wiki ( $M = 10,000$ ), NIPS ( $M = 1,740$ ).
- Size of consistent sets:  $k = 4$ ; dictionary size:  $K = 400$ .



LDA  $\approx$  OSDL  $\gg$  LSA. Wiki  $\gg$  NIPS (larger corpus).  $\delta = 0.1$ .

# Illustrations: Odd One Out Puzzles - Beginner

( $\eta_1 = 0.005, \eta_2 = 0.02$ )

Consistent set of words				Odd one out
vote	election	candidate	voters	sony
church	orthodox	presbyterian	evangelical	buddhist
olympic	tournament	world	championship	acid
austria	german	austrian	vienna	scotland
devil	demon	hell	soul	boat
harry	potter	wizard	ron	manchester
superman	clark	luthor	kryptonite	division
magic	world	dark	creatures	microsoft

# Illustrations: Odd One Out Puzzles - Intermediate

( $\eta_1 = 0.1, \eta_2 = 0.2$ )

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
function	problems	polynomial	equation	physical

- Automated word puzzle generation framework, 3 pillars:
  - 1 simple document collection (corpus),
  - 2 topic model → topic dictionary,
  - 3 semantic similarity of word pairs.
- The generated puzzles can be:
  - of many different types,
  - domain-specific,
  - of different levels (beginners, ...).
- Novel application of group-structured dictionaries.

# Acknowledgments

The research has been supported by the 'European Robotic Surgery' EC FP7 grant (no.: 288233).

Nemzeti Fejlesztési Ügynökség  
[www.ujszechenyiterv.gov.hu](http://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).

Thank you for the attention!



-  Alsumait, L., Barbará, D., Gentle, J., and Domeniconi, C. (2009).  
Topic significance ranking of LDA generative models.  
In *ECML PKDD '09*, pages 67–82.
-  Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012).  
Optimization with sparsity-inducing penalties.  
*Found. Trend. Mach. Learn.*, 4(1):1–106.
-  Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. (2010).  
Model-based compressive sensing.  
*IEEE T. Inform. Theory*, 56:1982–2001.
-  Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).  
Latent Dirichlet allocation.  
*J. Mach. Learn. Res.*, 3:993–1022.
-  Carter, P. (2005).  
*IQ and Psychometric Test Workbook*.



-  Colton, S. (2002).  
Automated puzzle generation.  
*In AISB'02, Imperial College, London.*
-  Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).  
Indexing by latent semantic analysis.  
*J. Am. Soc. Inform. Sci.*, 41(6):391–407.
-  Gabrilovich, E. and Markovitch, S. (2009).  
Wikipedia-based semantic interpretation for natural language processing.  
*J. Artif. Intell. Res.*, 34:443–498.
-  Jungnickel, D. (2007).  
*Graphs, Networks and Algorithms.*  
Springer; 3rd edition.
-  Pintér, B., Vörös, G., Szabó, Z., and Lőrincz, A. (2012).

Automated word puzzle generation using topic models and semantic relatedness measures.

In *Annales Univ. Sci. Budapest., Sect. Comp.*, volume 36, pages 299–322.

[http://ac.inf.elte.hu/Vol\\_036\\_2012/299\\_36.pdf](http://ac.inf.elte.hu/Vol_036_2012/299_36.pdf).



Szabó, Z., Póczos, B., and Lőrincz, A. (2011).

Online group-structured dictionary learning.

In *CVPR*, pages 2865–2872.



Verguts, T. and Boeck, P. D. (2000).

A Rasch model for detecting learning while solving an intelligence test.

*Appl. Psych. Meas.*, 24(2):151–162.