# Subspace Representation for Natural Language Processing

Erica Kido Shimomoto[1], François Portet[2], Kazuhiro Fukui[1]

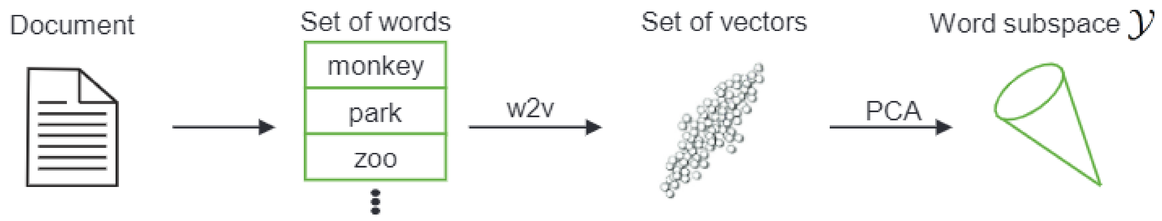
[1] *Graduate School of Systems and Information Engineering, University of Tsukuba, Japan*

[2] *University of Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*

## (1) Motivation and objective

- Exploring the applications of the **subspace-based methods** on **Word Embeddings**.
- Word Embeddings: Arithmetic and distance calculation between two word vectors -> Semantic relationship.
- **Word subspace [1]:**
  - Modeled using the Principal Components Analysis -> Low computational cost;
  - No restriction to the number of words;
  - Basis vectors = **Main hidden topics**;
  - Texts can be easily compared based on **subspace similarity**.
- Already applied to text summarization [2] and content generation [3];
- **Our goal**: Topic Classification and Sentiment Analysis.

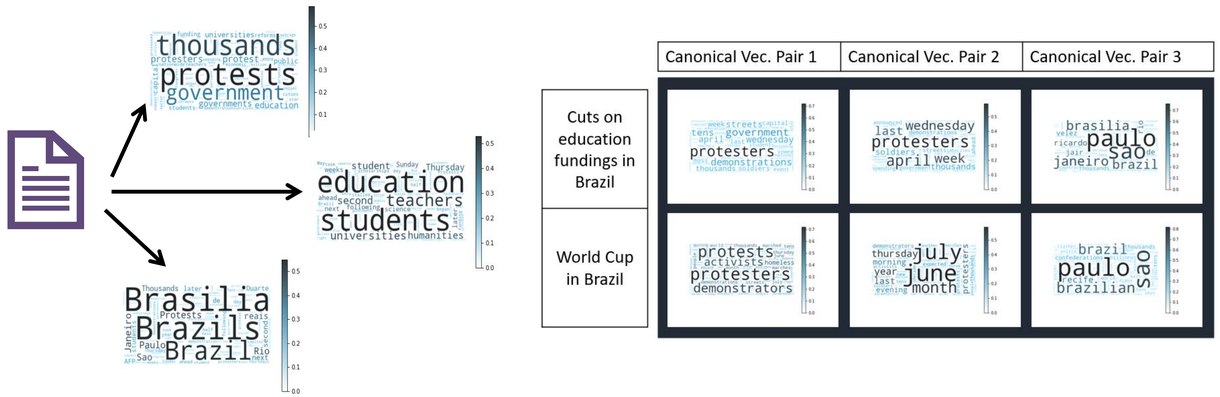
## (2) Word Subspace Modeling




## (3) Text Analysis based on the subspace representation

- **Word Importance Score:** Measures how relevant a word $w_k$ is with regards to a hidden topic $\boldsymbol{\phi}_i$.

$$I(\boldsymbol{x}_k, \boldsymbol{\phi}_i) = \boldsymbol{x}_k{}^{\mathsf{T}} \boldsymbol{\phi}_i$$




## (4) Topic classification

- Used the Mutual Subspace Method [4];

| Method | Text Model | R8 P | R8 R | R8 F1 | 20n P | 20n R | 20n F1 |
|---|---|---|---|---|---|---|---|
| MSM | u-WSub | 95.00 | 94.83 | 94.81 | **74.93** | **74.73** | **74.65** |
| MSM | WSub | 95.51 | 95.29 | 95.34 | 74.32 | 73.86 | 73.77 |
| SVM | PCA | 83.83 | 83.42 | 83.41 | 55.43 | 54.67 | 54.77 |
| SVM | p-mean | 96.69 | 96.67 | 96.65 | 72.20 | 71.65 | 71.79 |
| SVM | DCT | 96.98 | 96.98 | 96.94 | 72.20 | 71.58 | 71.73 |
| SVM | EigenSent | **97.18** | **97.13** | **97.14** | 72.24 | 71.62 | 71.78 |

- For more details:

E Shimomoto, L Souza, B Gatto, K Fukui, **Text Classification based on Word Subspace with Term-Frequency** (IJCNN18).

## (5) Sentiment Analysis

- Challenges:
  - Lack of sentiment information in word embeddings;
  - Same sentiment class can have texts of different topics.
- Proposed solutions:
  - Add discriminative power by using OMSM [5];
  - Represent sentiment class on the Grassmann manifold by using GSM and GOSM.

| Word Emb. | Method | Text Model | Movie Review | SST-2 |
|---|---|---|---|---|
| w2v | MSM | WSub | 76.45 | 75.53 |
| | GOSM | WSub | **84.25** | 72.91 |
| | LogReg | PCA | 65.74 | 71.94 |
| | LogReg | p-mean | 76.30 | 79.90 |
| | LogReg | DCT | 77.10 | **81.00** |
| GloVe | MSM | WSub | 76.80 | 77.12 |
| | GOSM | WSub | **85.75** | 67.80 |
| | LogReg | PCA | 63.43 | 50.58 |
| | LogReg | p-mean | 77.10 | 80.20 |
| | LogReg | DCT | 77.05 | 79.63 |
| | LogReg | WR | - | 82.20 |
| | LogReg | GEM | 78.80 | **83.60** |

## (6) Conclusions and Future Work

- We presented the Word Subspace to model texts based on Word Embeddings;
- We demonstrated its efficiency in the tasks of topic classification and sentiment analysis.
- Include word order -> RTW, SSA, SFA.

## (7) References

[1] Shimomoto, E. K., Souza, L. S., Gatto, B. B., and Fukui, K., “Text classification based on word subspace with term-frequency," in 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, IEEE, 2018.

[2] Gong, H., Sakakini, T., Bhat, S., Xiong, J.: Document similarity for texts of varying lengths via hidden topics. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2341-2351. Association for Computational Linguistics, Melbourne, Australia (2018).

[3] Shimomoto, Erica K., et al. "News2meme: An Automatic Content Generator from News Based on Word Subspaces from Text and Image." 2019 16th International Conference on Machine Vision Applications (MVA). IEEE, 2019.

[4] Fukui, K. and Maki, A., “Difference subspace and its generalization for subspace-based methods.” IEEE transactions on pattern analysis and machine intelligence. 2015.

[5] 河原智一, 西山正志, and 山口修. "直交相互部分空間法を用いた顔認識." 情報処理学会研究報告コンピュータビジョンとイメージメディア (CVIM) 2005.112 (2005-CVIM-151) (2005): 17-24.