Word Embedding en chinois

Réalisation par :
Lu Chenxin
Xu Jitao

Sous supervision de :
M. Stéphane Canu

PLAN

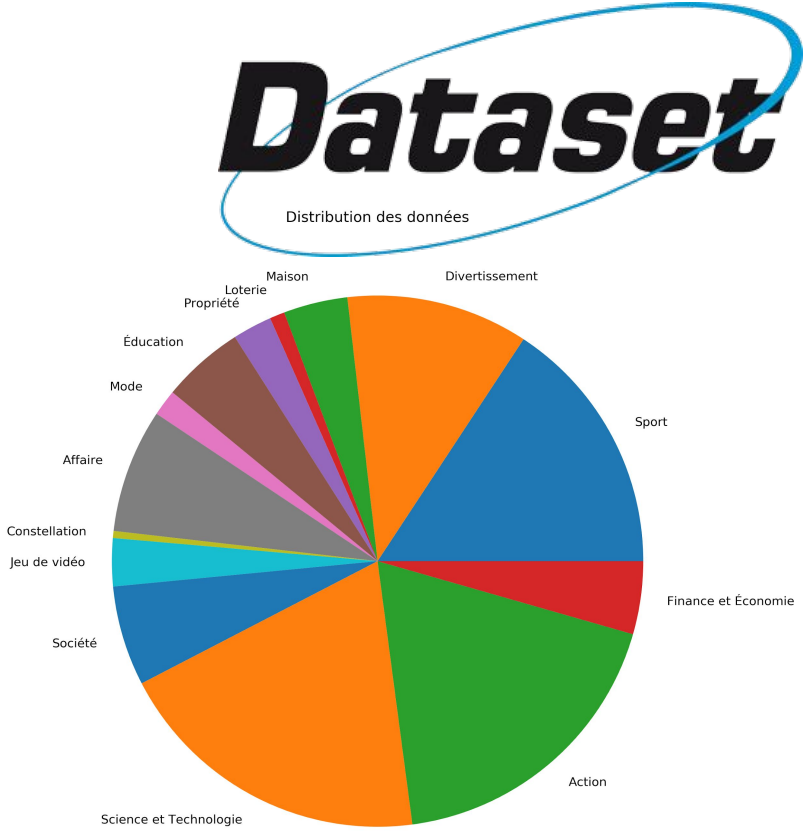
1. Introduction
2. Jeux de données
3. Modèles existants de Word Embedding
4. Evaluations de Word Embedding
5. Résultats et Conclusions
6. Bibliographie

1. Introduction

- **Contexte** : Les jeux de données (les corpus) en chinois
- **Objectifs** : Comparer les méthodes de Word Embedding existantes
  - Trouver des données textuelles en chinois utilisables
  - Identifier les modèles existants
  - Proposer des bons critères de comparaison
  - Choisir une tâche en aval pour comparer les Word Embedding

2. Jeux de données

- Sport : 15.74%
- Divertissement : 11.08%
- Maison : 3.90%
- Loterie : 0.90%
- Propriété : 2.40%
- Éducation : 5.02%
- Mode : 1.60%
- Affaire : 7.54%
- Constellation : 0.43%
- Jeu de vidéo : 2.92%
- Société : 6.08%
- Science et Technologie : 19.49%
- Action : 18.47%
- Finance et Économie : 4.44%



2. Jeux de données

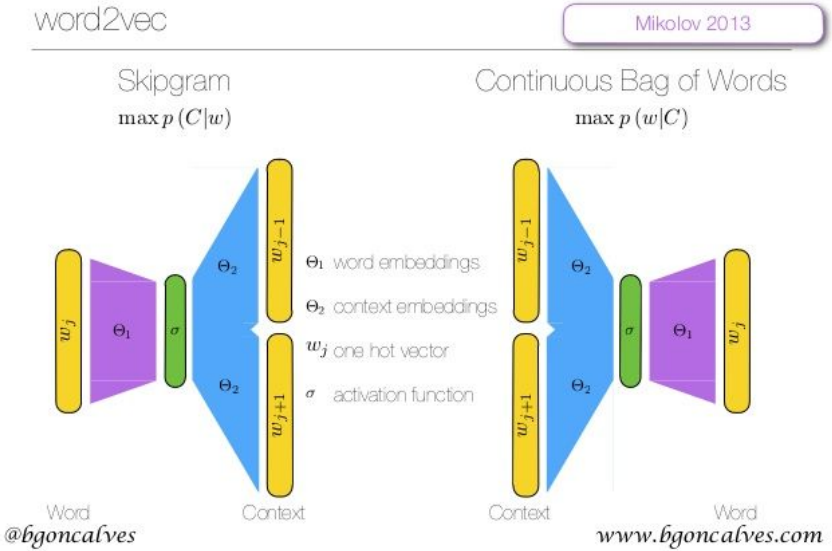
- Sources : un corpus de nouvelles de THUCTC
- Deux fichiers séparés : Les jeux de données d'apprentissage et de test
- Un exemple contient une nouvelle
- Étiquettes : Le type de nouvelles
- Nombre de classes : 11 (Nous avons éliminé 3 classes qui ont moins de 10000 exemples)
- Tailles :
    - Le jeu de données d'apprentissage : 109989
    - Le jeu de données de test : 103369

3. Modèles existants de Word Embedding

1. Word2Vec
2. GloVe
3. FastText
4. ELMo

3.1 Word2Vec

- Skip-gram et CBOW
- Hierarchical Softmax
- Negative Sampling



3.2 GloVe

- La matrice de co-occurrence
- La relation linéaire entre les mots

3.3 FastText

- N-gram
- Hierarchical Softmax
- Une méthode de classification de texte

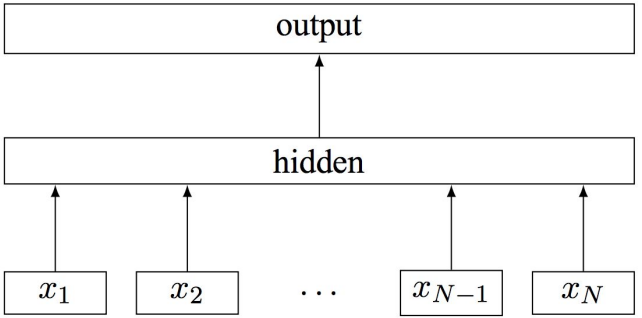


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

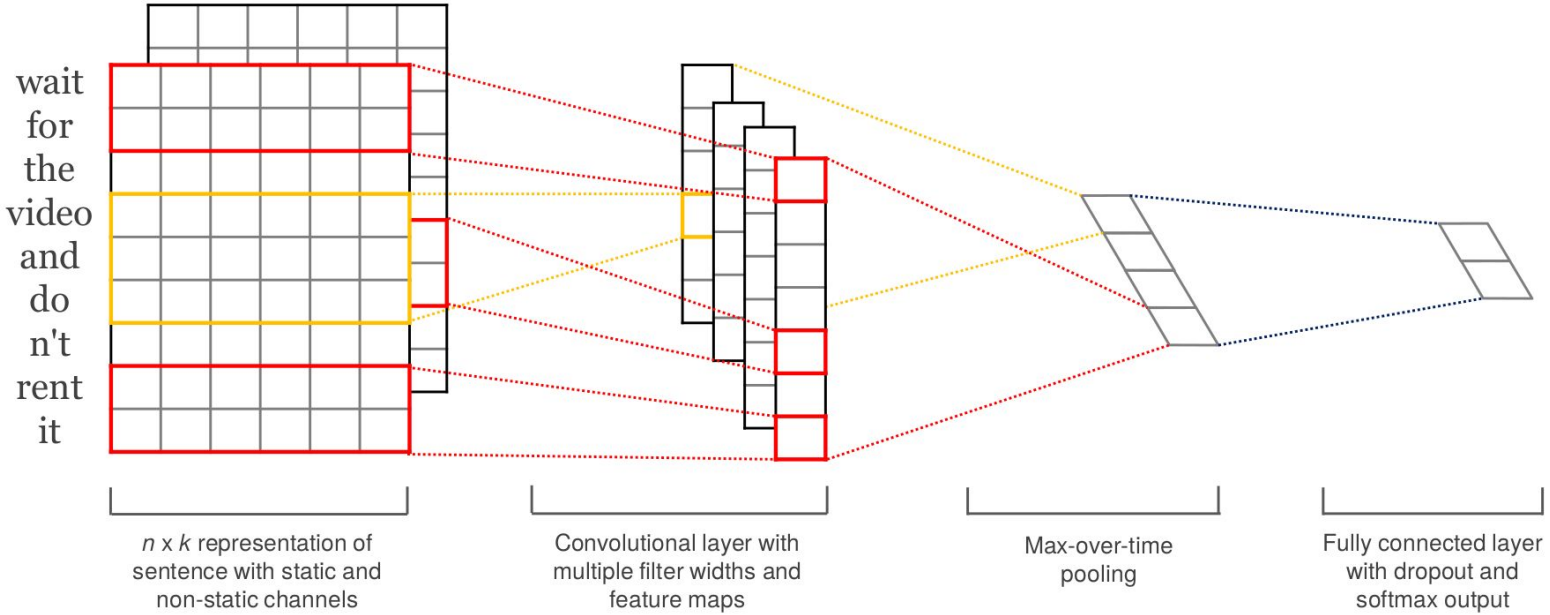
3.4 ELMo

- représentation contextualisée
- Bi-LSTM
- Une concaténation d’une représentation non-contextualisée et deux représentations contextualisées
- Deux couches de LSTM représentent différents caractéristiques : Sémantique et Syntaxique

4. Evaluations de Word Embedding

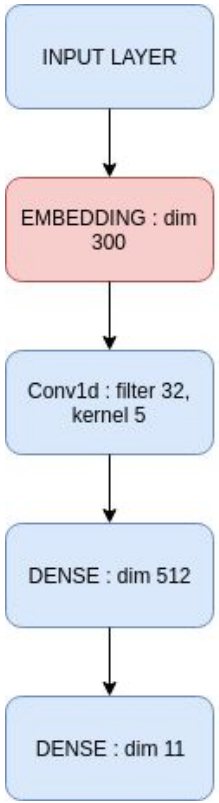
- Évaluation non-contextualisée
  - Relations Morphologiques
    - Duplication : "tian"(jour) → "tian tian"(tous les jours)
    - Semi-fixation : "yi"(un) → "di - yi" (première)
  - Relations Sémantiques : par exemple, les capitales de pays
- Évaluation contextualisée
  - La classification de texte : TextCNN

4.2 TextCNN

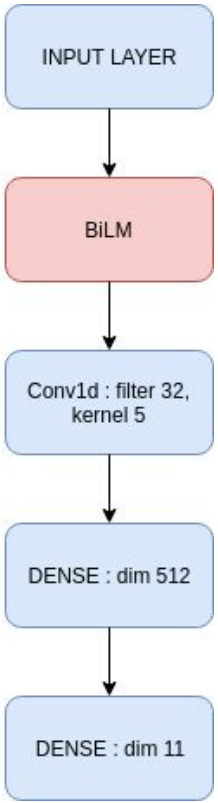

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

4.2 TextCNN

CNN Pipeline



CNN for ELMo

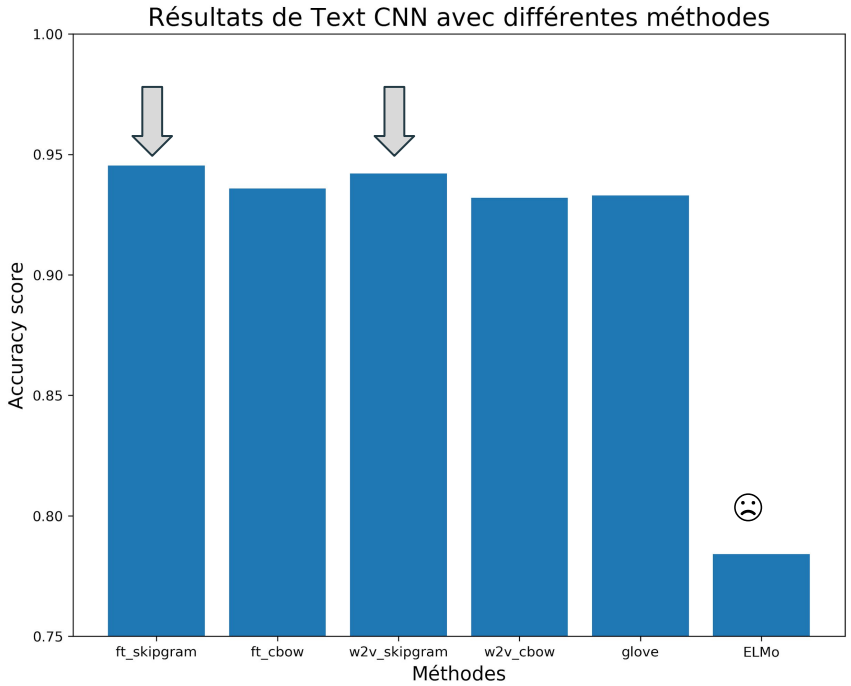


5. Résultats et Conclusions

Méthode	Geography	Nature	History	People	**Semantic**
FastText SkipGram	**0.359**	0.14	0.0	0.207	**0.276**
FastText CBOW	0.222	0.133	**0.007**	0.102	0.178
Word2Vec SkipGram	0.325	0.173	0.0	0.213	0.264
Word2Vec CBOW	0.179	**0.18**	0.0	**0.218**	0.179
GloVe	0.184	0.105	**0.007**	0.202	0.163

Méthode	A	AB	Prefix	Suffix	**Morphological**
FastText SkipGram	0.013	0.018	0.077	0.18	0.081
FastText CBOW	0.011	0.015	**0.129**	**0.25**	**0.115**
Word2Vec SkipGram	0.018	0.029	0.048	0.095	0.052
Word2Vec CBOW	**0.023**	**0.043**	0.111	0.127	0.082
GloVe	0.004	0.001	0.019	0.018	0.012

5. Résultats et Conclusions



5. Résultats et Conclusions

- Sémantique → Skip Gram
- Morphologique → CBOW
- FastText → Meilleur!!!
- ELMo → À explorer
- Transformer → BERT...

6. Bibliographie

[1] Piotr BOJANOWSKI et al. "Enriching Word Vectors with Subword Information". In : *CoRR* abs/1607.04606 (2016). arXiv : 1607.04606. URL : http://arxiv.org/abs/1607.04606.

[2] Armand JOULIN et al. "Bag of Tricks for Efficient Text Classification". In : *CoRR* abs/1607.01759 (2016). arXiv : 1607.01759. URL : http://arxiv.org/abs/1607.01759.

[3] Yoon KIM. "Convolutional Neural Networks for Sentence Classification". In : *CoRR* abs/1408.5882 (2014). arXiv : 1408.5882. URL : http://arxiv.org/abs/1408.5882.

[4] Shen LI et al. "Analogical Reasoning on Chinese Morphological and Semantic Relations". In : *CoRR* abs/1805.06504 (2018). arXiv : 1805.06504. URL : http://arxiv.org/abs/1805.06504.

[5] Tomas MIKOLOV, Kai CHEN et al. "Efficient Estimation of Word Representations in Vector Space". In : *CoRR* abs/1301.3781 (2013). arXiv : 1301.3781. URL : http://arxiv.org/abs/1301.3781.

[6] Tomas MIKOLOV, Ilya SUTSKEVER et al. "Distributed Representations of Words and Phrases and their Compositionality". In : *CoRR* abs/1310.4546 (2013). arXiv : 1310.4546. URL : http://arxiv.org/abs/1310.4546.

[7] Jeffrey PENNINGTON, Richard SOCHER et Christopher D. MANNING. "GloVe : Global Vectors for Word Representation". In : *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, p. 1532–1543. URL : http://www.aclweb.org/anthology/D14-1162.

[8] Matthew E. PETERS et al. "Deep contextualized word representations". In : *CoRR* abs/1802.05365 (2018). arXiv : 1802.05365. URL : http://arxiv.org/abs/1802.05365.

[9] Ashish VASWANI et al. "Attention Is All You Need". In : *CoRR* abs/1706.03762 (2017). arXiv : 1706.03762. URL : http://arxiv.org/abs/1706.03762.


We need to go deeper