Juin 2020

# On the evaluation of retrofitting for supervised short-text classification

Kaoutar GHAZI [a], Andon TCHECHMEDJIEV [a], Sébastien HARISPE [a],
Nicolas SUTTON-CHARANI [a] and Gildas TAGNY NGOMPÉ [b]
[a] *EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Alès, Alès, France*
[b] *ESII, Laverune - FRANCE*

**Abstract.** Current NLP systems heavily rely on embedding techniques that are used to automatically encode relevant information about linguistic entities of interest (e.g., words, sentences) into latent spaces. These embeddings are currently the cornerstone of the best machine learning systems used in a large variety of problems such as text classification. Interestingly, state-of-the-art embeddings are commonly only computed ~~only~~ using large corpora ~~analysis~~, and generally do not use additional knowledge expressed into ~~existing~~ established knowledge resources ~~representations~~ (e.g. WordNet). In this paper, we empirically study if retrofitting, a class of techniques used to update word vectors in a way that takes into account knowledge expressed in knowledge resources, is beneficial for ~~supervised~~ short text classification. To this aim, we compare the performances of several state-of-the-art classification techniques with or without retrofitting on a selection of benchmark datasets. Our results show that...

**Keywords.** supervised text classification, word embeddings, retrofitting

## Introduction

Embedding techniques are the cornerstone of numerous state-of-the-art NLP systems; they enable to automatically encode relevant information about linguistic entities of interest (e.g., words, sentences, documents) into latent spaces in order to obtain high quality representations that will further be used to solve complex tasks. Such techniques have proven to be critical for designing efficient systems in text classification [?], question answering [?] or information extraction [?] to mention a few.

Neural ~~N~~network~~-based~~ architectures, particularly recurrent neural networks (RNN) or transformers (attention mechanism) are now *de facto* approaches to computing embeddings, as illustrated by the broad variety of language models of increasing complexity and efficiency that have been published in recent years (e.g. RoBERTa, GPT-3). These approaches ~~only~~ rely on the surface analysis of large corpora composed of billions of words, and do not use additional knowledge expressed into ~~existing~~ established knowledge resources ~~representations~~ (e.g. WordNet). Despite recent successes, there is only so much that can be learned from a surface analysis of text and embedding models capture very superficial knowledge about meaning [16]. One way of integrating structured *a pri-*

C'était déjà Juin sur la page précédente

*ori* knowledge, is to apply *retrofitting*, a class of techniques used to update word vectors in a way that takes into account knowledge expressed in knowledge resources. Despite the ~~interesting~~ promising results obtained by retrofitting techniques, the study of hybrid embedding approaches mixing corpora and knowledge representations is still relatively marginal, especially in context of specific tasks. This paper aims at investigating the relevance of retrofitted word embeddings in the context of supervised short-text classification, especially when compared to state of the art contextualised language models. We compare the performance of several pre-trained word embedding models with and without retrofitting for short-text classification. We explore several retrofitting approaches and use word vectors as features with both a classical machine-learning pipeline and a more state of the art bi-LSTM encoder. We further compare to two transformer baselines, where the transformers are directly used for classification.

The paper is organized as follows: Section 1 briefly ~~briefly~~ present the two tow most common retrofitting models; Section 2 presents the protocol used in our experimental setting as well as the obtained results ~~we obtain~~. Section 3 discusses those results and offers additional observations that question the benefit of current retrofitting approaches for the studied task.

## 1. Retroffiting embeddings in NLP

State-of-the-art word embedding techniques solely based on corpora analysis are framed on the distributional hypothesis stating that words occurring in similar contexts tend to be semantically close [31]. This hypothesis, made popular through ~~the~~ Firth's idea ~~belonging to Firth~~ (1957) [32]: "*You shall know a word by the company it keeps*", is one of the main tenets of statistical semantics. By definition such approaches do not ~~have access to~~ use lexical or conceptual relationships that could be important to characterize words' semantic~~s~~, e.g. some approaches will similarly represent synonyms and antonyms [12]. To address this limitation, a class of approaches denoted *Retrofitting* aims at incorporating *a priori* knowledge from external resources ~~while computing~~ to refine word embeddings, e.g. lexicons, ontologies, domain-specific datasets expressing semantic knowledge such as paraphrase.

A mon avis la phrase précédente, n'exprime pas une limitation mais juste le fait que ces relations ne sont pas utilisées. Il faut probalement dire en quoi c'est limitant, dans quelles situations.

~~A general approach adopted in the literature~~ The use of external data or knowledge generally requires retraining the model used to compute the embeddings (considered as a subset of the model's parameters). In this case retrofitting can be seen as a post-processing step aiming at updating pre-trained word embeddings in order to induce a refined vector space with desired properties encoded in the external resource ~~in use~~ (such as FrameNet, PPDB and WordNet). Indeed, in addition to observed words contexts (i.e. surroundings), semantic lexicons resources such as sets of words labelled with semantic relation (e.g. hyperonymy, hyponymy) can be used. To this aim, a specific objective function is for instance defined to ~~consider~~ learn both the distribution of words and their lexical (resp. conceptual) relationships either jointly [14,15,17,18] or ~~alternatively~~ separately [9]. Another strategy proposes to consider independent representations f~~o~~rom corpora and knowledge representations to later combine them, e.g. Goikoetxea et al. defined a method that independently learns word representations from WordNet, to later combined them with text-based representations (their embeddings→ les 2 sont des embeddings non?) [22]. Several contributions have been proposed to refine these general strategies, e.g. Vulić et al. have proposed an approach based on context analysis enabling to retrofit words that do not occur into lexicon [13] that occur into similar dependency-based contexts ; Yih et al. proposed to combine the use of a thesaurus to distinguish synonyms from antonyms [12].

that ... that (c'est bien enchainé comme ça ou bien c'est l'un ou l'autre?)

These approaches have traditionally been proposed for static word representations, i.e. a single representation is associated to a word (token). Recently, contextualized text-embedding models have been proposed to deal with issues induced by polysemy [19,20]. In this case, a context-specific representation~~s~~ of a word is obtained depending on its meaning in each sentence. Recent retrofitting techniques are designed for these contextualized embeddings, e.g. Shi et al. propose to consider prior knowledge about paraphrases to improve context-specific representations.

Several studies have stressed the benefits of retrofitting on several NLP tasks such as sentiment analysis, relationship and text classifications [9,10,11,12,13,14,15,17]. These studies only contain limited comparisons with state-of-art text- embeddings, in particular ~~wrt to~~ for short-text classification.

## 2. Evaluation protocol and Results

In this ~~paper~~ section, we ~~focus on~~ evaluat~~ing~~e the ~~interest~~ benefit of retrofitting approach for short-text classification. We consider the “retrofitting” technique of Faruqui et al. [9] and the “counterfitting” strategy of [10]. Mrkšić et al. (→ il faudrait l’introduire en section 1). il faut justifier le choix de ces 2 stratégies. ce serait peut-être bien de les introduire de cette façon.

### *2.1. Datasets*

**Word Embeddings:** we consider the~~se~~ following 300-dimensional word vectors : (i) Paragram [22], learned from the paraphrase database PPDB (ii) Glove [8], learned from Wikipedia and Common Crawl data (iii) MUSE, a fast-text embedding learned from Wikipedia [1] (iv) contextualized word embeddings models : Flair [29] and RoBERTa [23]. For each word embedding~~s~~, except the contextualized ones, we consider three versions: baseline, retrofitted and counterfitted.

**Semantic lexicons:** we consider the state-of-the-art lexical resources (i) PPDB [25], a collection of English paraphrases from which a set of antonyms and synonyms are identified [9], [10] (ii) FrameNet [26], a collection of frames grouping related words (iii) WordNet [27], a linguistic ontology grouping words from similar lexical category (verb, noun, etc.) into synsets (synonym sets) which are interconnected by semantic relations, i.e. hyponymy. In this paper, we distinguish between WordNet+, that takes into consideration all semantic relations (synonyms, hypernyms and hyponyms), and WordNet$_{syn}$ that only consider synonym~~s~~ relations.

Peut-être ça peut couté moins cher en expérimentations en testant d'abort les meilleures configurations des articles des stratégies choisies. Par exemple, le papier sur le counter-fitting (Nikola Mrksic et al, 2016) la combinaison "WordNet− and PPDB+" donne les meilleurs résultats

**Evaluation benchmarks**

The evaluations were performed on the following benchmark datasets:

[1] https://github.com/facebookresearch/MUSE

- *HuffPost headlines*~~,~~ [28] [2] ~~a set of:~~ 200849 headlines published on HuffPost from the year 2012 to 2018. Each news headline belongs to one of the 41 possible categories.
- *Product Listing On Amazon India*~~,~~ [3] ~~a set of:~~ 27375 product titles from Amazon India of the year 2019. We keep product belonging to only one of the 9 distinct categories available and we drop redundant records.
- *Amazon Consumer Reviews of Amazon Products* [4] ~~a set of:~~ 28332 customer reviews for Amazon products categorized into 60 categories.

### 2.2. Evaluation Process

~~We use three machine learning models : ridge classifier, RandomForestClassifier and XGBClassifier.~~ We consider two different modelling contexts: shallow machine learning where the sequence of words is not taken into account but computation times are small and deep learning where the order of words apparition in text is used for the classification task, with necessarily higher computation times. In the first context, three models are compared: the ridge classifier, random forest and XGBoost. ~~and~~ Flair is used as LSTM baseline. For each model, we apply a grid search to find its best hyper-parameter values then we run a 10-folds cross validation on models optimally ~~configurated~~ tuned. Data used for the grid-search~~ing~~ and the cross validation are independent splits extracted from the ~~benchmark~~ considered dataset. Words embeddings (baseline, retrofitted or counter-fitted) are given in input of each model. We keep only the best shallow machine learning classifier, in term of accuracy ~~of classification~~, and compare it with Flair model~~, an LSTM neural network for text classification~~.

Dire quelle méthode de pooling a été utilisée sur les word embeddings pour représenter les phrases en entrées des algorithmes statistiques

Généralement, c'est suffisant et moins cher d'utiliser le même simple train-test d'un autre article pour les mêmes données. En plus, un train-test permet de recommander des valeurs d'hyperparamètres

### 2.3. Evaluation Results

In Table 1 we report performances, in term of accuracy, of ~~the best machine learning classifier and the deep learning classifier with LSTM architecture (Flair model) on the benchmark considered~~ all models on the three benchmark datasets. Results show that the Glove words embedding (the baseline) are relevant for the task of text classification with ridge classifier and do not need to be retrofitted (i.e. vectors are sufficiently significant). We also observe that contextualized words embedding (BERT model) is much more meaningful and overall, it provides best results. However, this embedding, if not already pre-trained, requires a huge amount of resources to be useful.

le ridge est le seul présenté parmis les 3 modèles traditionnels.

peut-être pas "pertinent" ou " sufficiently significant", vu qu'il ne dépasse même ps les 50%, mais c'est sûr que le retrofitting tel qu'il a été expérimenté, n'améliore pas ses résultats

peut-être l'expérimentation du retrofitting de BERT aurait été suffisante ou au moins aussi intéressante à réaliser.

The LSTM architecture is not powerful for the classification of the dataset considered because of the none uniform distribution of records across class labels; e.g. more than 16000 records for some classes (overfitting issue) and less than 2000 for others (underfitting issue) in HuffPost headlines dataset. However, we observe that deep learning model take advantage from anthologies knowledge injected into words embeddings. Table 2 gives an illustration of this for the HuffPost Headlines dataset restrained in 3 classes (Travel, Style & Beauty, Parenting) with almost 10000 records for each.

[2] https://www.kaggle.com

[3] https://data.world/promptcloud/product-listing-on-amazon-india

[4] https://data.world/

| Paragram | | | Glove | | | MUSE | | |
|---|---|---|---|---|---|---|---|---|
| Baseline | Retrofit | Counterfit | Baseline | Retrofit | Counterfit | Baseline | Retrofit | Counterfit |
| | | | | | | | | |

**Table 1.** Accuracy of classification for LSTM-NN for Huffpost headlines (3 class labels)

| Embeddings | Semantic Lexicon | HuffPost Headlines | | Product Amazon India | | Consumer Reviews | |
|---|---|---|---|---|---|---|---|
| | | RidgeC | LSTM | RidgeC | LSTM | RidgeC | LSTM |
| Paragram | | 42.77 | | 36.67 | | 85.56 | 50.68/43.69 |
| Paragram Retrofitted | PPDB | 42.80 | 66.78 | 36.63 | | 85.37 | 42.51 |
| | FrameNet | 42.57 | | 36.39 | | 85.04 | 51.30 |
| | WordNet$_{syn}$ | 42.63 | | 36.56 | | 85.23 | 38.75 |
| | WordNet+ | 42.38 | | 36.64 | | 85.12 | 44.56 |
| Paragram Counterfitted | PPDB&WordNet | 42.57 | | 35.97 | | 85.49 | |
| Glove | | 45.28 | 65.06 | 36.94 | | 85.99 | 40.74 |
| Glove Retrofitted | PPDB | 45.40 | | 36.84 | | 85.94 | 40.32 |
| | FrameNet | 44.89 | | 36.87 | | 85.07 | 38.29 |
| | WordNet$_{syn}$ | 44.69 | 63.80 | 36.54 | | 85.72 | 42.61 |
| | WordNet+ | 44.52 | | 36.56 | | 85.65 | 43.60 |
| Glove Counterfitted | PPDB&WordNet | 44.32 | | 36.63 | | 85.76 | |
| MUSE | | 44.80 | | 36.26 | | 84.58 | 43.44/43.82 |
| MUSE Retrofitted | PPDB | 45.20 | | 36.25 | | 84.45 | 40.06 |
| | FrameNet | 44.43 | | 35.87 | | 84.02 | 44.59 |
| | WordNet$_{syn}$ | 44.27 | | 36.40 | | 84.03 | 40.9 |
| | WordNet+ | 44.20 | | 36.10 | | 83.91 | 42.10 |
| MUSE Counterfitted | PPDB&WordNet | 44.07 | | 36.01 | | 84.61 | |
| Flair | | 41.81 | | 34.69 | | 86.62 | |
| RoBERTa | | 53.48 | | 39.24 | | 86.92 | |

**Table 2.** Accuracy of classification for the ridge classifier and LSTM-NN with the baseline word embeddings then with the retrofitted, counterfitted and contextualized word embeddings

## 3. Discussion

Ce paragraphe me parait bien pour introduire les 2 stratégies expérimentées dans la section 1. Retroffiting embeddings in NLP

According to the state of the art, both retrofitting techniques were evaluated on words similarity estimation task. As a results, the standard retrofitting technique improves words embeddings on the three following datasets : WS-353 [2], a dataset of 353 English words, RG-65 [4] a dataset of 65 pairs of nouns, and MEN [3], a dataset of 3000 word pairs randomly selected from words occurring at least 700 times in ukWaC and Wackypedia corpora [5]. In turn, the counterffing technique improves words embeddings on SimLex-999 [1], a dataset of 999 word pairs for which the similarity is measured without considering the relatedness or the association between words. Using this later dataset, Mrkšić et al. [10] show that there technique of counterfitting (CT) is almost 5% better than the standard retrofitting technique (SRT) for this task.

Toujours pour la section 1. Retroffiting embeddings in NLP, Ces paragraphes justifient assez bien l'objectif du travail

Although both techniques have also proven their efficiency on some extrinsic tasks such as sentiment analysis (for CT [9]) and dialogue state tracking (for SRT [10]), none

[5] https://staff.fnwi.uva.nl/e.bruni/MEN

comparison of these techniques were studied at that level. However, this still not enough to conclude which technique is better for short text classification task or even if the retrofitting approach is recommanded in that context.

When dealing with text classification, it is obvious that the more significant word embeddings are, the more accurate classification of texts is expectable. Therefore, we propose in this paper an evaluation of techniques proposed in the literature for improving words sens representations in order to underline their efficiency in increasing texts classifiers accuracy.

Recently, contextuel word embeddings provide meaningful representation of words according to the current context in which they occur. We questioned how beneficial are these embeddings than retrofitted ones and what about using contextuel retrofitted embeddings for text classification.

According to our evaluation results presented in Table 1, we observe that retrofitting techniques improve word embeddings by x% in average.

## 4. Conclusion

Les noms de journaux et de conférences, et les numéro de pages et de volumes n'apparaissent pas pour de nombreux articles.

## References

[1] Hill, Felix and Reichart, Roi and Korhonen, Anna. Simlex-999: Evaluating semantic models with (genuine) similarity estimation, 2015.

[2] Finkelstein, Lev and Gabrilovich, Evgeniy and Matias, Yossi and Rivlin, Ehud and Solan, Zach and Wolfman, Gadi and Ruppin, Eytan. Placing search in context: The concept revisited, 2001.

[3] Bruni, Elia and Boleda, Gemma and Baroni, Marco and Tran, Nam-Khanh. Distributional semantics in technicolor, 2012.

[4] Rubenstein, Herbert and Goodenough, John B. Contextual correlates of synonymy, 1965.

[5] Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey. Efficient estimation of word representations in vector space, 2013.

[6] Salton, Gerard and Wong, Anita and Yang, Chung-Shu. A vector space model for automatic indexing. In: ACM New York, 1975.

[7] Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In : Advances in neural information processing systems, 2013.

[8] JeffreyPennington, RichardSocher and Manning, ChristopherD. Glove: Global vectors for word representation. In : Conference on Empirical Methods in Natural Language Processing, 2014.

[9] Faruqui, Manaal and Dodge, Jesse and Jauhar, Sujay K and Dyer, Chris and Hovy, Eduard and Smith, Noah A. Retrofitting word vectors to semantic lexicons. In : Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2014.

[10] Mrkšić, Nikola and Séaghdha, Diarmuid O and Thomson, Blaise and Gašić, Milica and Rojas-Barahona, Lina and Su, Pei-Hao and Vandyke, David and Wen, Tsung-Hsien and Young, Steve. Counter-fitting word vectors to linguistic constraints. In : Proceedings of NAACL-HLT, 2016.

[11] CHIU, Billy, BAKER, Simon, PALMER, Martha, et al. Enhancing biomedical word embeddings by retrofitting to verb clusters. In : Proceedings of the 18th BioNLP Workshop and Shared Task, 2019.

[12] YIH, Wen-tau, ZWEIG, Geoffrey, et PLATT, John C. Polarity inducing latent semantic analysis. In : Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.

[13] VULIĆ, Ivan, SCHWARTZ, Roy, RAPPOPORT, Ari, et al. Automatic selection of context configurations for improved class-specific word representations. arXiv preprint arXiv:1608.05528, 2016.

[14] YU, Mo et DREDZE, Mark. Improving lexical embeddings with semantic knowledge. In : Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014.

[15] BIAN, Jiang, GAO, Bin, et LIU, Tie-Yan. Knowledge-powered deep learning for word embedding. In : Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2014.

[16] Bender, Emily M., Koller, Alexander. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198. July 5 - 10, 2020.

[17] XU, Chang, BAI, Yalong, BIAN, Jiang, et al. Rc-net: A general framework for incorporating knowledge into word representations. In : Proceedings of the 23rd ACM international conference on conference on information and knowledge management, 2014.

[18] FRIED, Daniel et DUH, Kevin. Incorporating both distributional and relational semantics in word representations. arXiv preprint arXiv:1412.4369, 2014.

[19] PETERS, Matthew E., NEUMANN, Mark, IYYER, Mohit, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.

[20] DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[21] SHI, Weijia, CHEN, Muhao, ZHOU, Pei, et al. Retrofitting contextualized word embeddings with paraphrases. arXiv preprint arXiv:1909.09700, 2019.

[22] GOIKOETXEA, Josu, AGIRRE, Eneko, et SOROA, Aitor. Single or multiple? combining word representations independently learned from text and wordnet. In : Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[23] LIU, Yinhan, OTT, Myle, GOYAL, Naman, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

```
[1] @article{hill2015simlex,
  title={Simlex-999: Evaluating semantic models with (genuine) similarity estimation},
  author={Hill, Felix and Reichart, Roi and Korhonen, Anna},
  journal={Computational Linguistics},
  volume={41},
  number={4},
  pages={665--695},
  year={2015},
  publisher={MIT Press}
}
```

```
[2] @inproceedings{finkelstein2001placing,
  title={Placing search in context: The concept revisited},
  author={Finkelstein, Lev and Gabrilovich, Evgeniy and Matias, Yossi and Rivlin, Ehud and Solan, Zach and Wolfman, Gadi and Ruppin, Eytan},
  booktitle={Proceedings of the 10th international conference on World Wide Web},
  pages={406--414},
  year={2001}
}
```

```
[3] @inproceedings{bruni2012distributional,
  title={Distributional semantics in technicolor},
  author={Bruni, Elia and Boleda, Gemma and Baroni, Marco and Tran, Nam-Khanh},
  booktitle={Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)},
  pages={136--145},
  year={2012}
}
```

[24] ZHU, Yukun, KIROS, Ryan, ZEMEL, Rich, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In : Proceedings of the IEEE international conference on computer vision, 2015.

[25] GANITKEVITCH, Juri, VAN DURME, Benjamin, et CALLISON-BURCH, Chris. PPDB: The paraphrase database. In : Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013.

[26] BAKER, Collin F., FILLMORE, Charles J., et LOWE, John B. The berkeley framenet project. In : Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1998.

[27] MILLER, George A. WordNet: a lexical database for English. Communications of the ACM, 1995.

[28] Rishabh Misra. News category dataset. 10.13140/RG.2.2.20331.18729 June 2018.

[29] Akbik, Alan and Blythe, Duncan and Vollgraf, Roland. Contextual String Embeddings for Sequence, 2018.

[30] Ting-Yu Yen, Yang-Yin Lee, Hen-Hsen Huang, Hsin-Hsi Chen. That Makes Sense: Joint Sense Retrofitting from Contextual and Ontological Information, 2018.

[31] Harris, Zellig S. Distributional structure. Papers in structural and transformational linguistics. 1970.

[32] Firth, John R. A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis. 1957