4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15076

Text mining methods for innovation studies: limits and future perspectives

Pietro Cruciata, Davide Pulizzotto, Catherine Beaudry
Polytechnique Montréal, Canada.

Abstract

This study offers alternative and promising approaches to word count methods, largely used to develop innovation indicators from unstructured text. We propose a method based on Information Retrieval (IR) and word-embedding models to tackle the semantic ellipsis, one of the main issues of word count methods. We test our IR model by investigating the concept of collaboration and comparing our approach with a baseline corresponding to the keyword search. To ensure the best performances, we use several ways to represent queries and documents in a vector space and three pre-trained word-embedding models. The results prove that our approach can alleviate the semantic ellipsis problem. Indeed, the IR model developed outperforms the classical keyword search in terms of F1-score and Recall. Moreover, we create a combined method that achieves the highest F1-score. These preliminary results can facilitate the creation of reliable innovation indicators from unstructured textual data substituting or complementing survey-based questionnaires.

Keywords: *Text mining, Natural Language Processing, Information Retrieval, innovation measures.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0
Editorial Universitat Politècnica de València

1. Introduction

The interest to analyze and comprehend innovation dynamics increased since technological progress became the main driver of economic growth, augmenting the need for researchers of public databases and questionnaire-based surveys as a source of data for their quantitative studies. Nevertheless, these sources of information have many weaknesses. Public databases are often incomplete or not specific whereas questionnaire-based surveys (especially large-scale as the biennial European CIS or the annual MIP) lack regional granularity, coverage, timeliness, and furthermore, they are costly (Axenbeck & Breithaupt, 2021). For all of these reasons traditional innovation indicators rarely provide the full picture (Kinne & Lenz, 2021)

Alternative or complementary to these sources is web-based unstructured textual data. Indeed, the rising amount of data available in the form of digitalized text opened up new possibilities for researchers. Although it seemed difficult to measure "signals" of innovation dynamics through corporate websites or other web sources, researchers on innovation and technology management have obtained good results by building new indicators with large amounts of texts. For example, Arora et al. (2013) built five descriptive variables by analyzing a sample of small and medium-sized high technology graphene firms in the US, UK, and China based on keyword analysis of their webpages. Gök et .al. (2015) created web indicators of R&D activities by extracting the keywords from companies' websites. Their study proved that R&D activities captured through the web indicators were significantly more numerous, compared to the R&D activities documented in the other sources. Libaers, et al. (2016) harnessed the data of the companies' websites to develop a taxonomy that identified strategies used by small firms to commercialize their innovations. They analyzed the content of firms' websites to extract the keywords related to possible strategies used by companies. Blazquez & Domenech (2018) used web-based variables built with keywords to predict firm export orientation. Héroux-Vaillancourt et al. (2020) built innovation indicators of four core concepts (R&D, IP protection, collaboration and external financing) from the complete texts of 79 corporate websites of Canadian nanotechnology and advanced materials firms using keywords frequency analysis.

As highlighted in the few examples above, most of the indicators created on textual data are based on keyword search and keyword frequency with several weighting schemes, such as TF-IDF. However, these indicators have two important drawbacks when it comes to analyzing concepts in texts: polysemy of words, a semantic phenomenon that illustrates the relationship between one word and multiple meanings (e.g., river "bank" vs. "bank" as a financial institution) and semantic ellipsis, which refers to the emerging of concepts in sentences where standard words referring to it have been omitted (e.g., the concept of "collaboration" can be expressed by using a combination of words such as "joint venture", "work with", "join forces" etc.). There are two tasks in Natural Language Processing (NLP) that apply advanced methodologies to solve these two issues. The first task is Word Sense Disambiguation (WSD) which refers to the problem of determining what is the word's meaning in a particular context. The other task is Information Retrieval (IR) whose goal is to search through documents to retrieve the best answer to a query.

In our research, we tackle the problem of the semantic ellipsis by developing a word-embedding-based IR model. Considering that collaboration is one of the main innovation indicators, we chose this concept to test our approach.

To the best of our knowledge, this is the first work that uses pretrained neural networks to preprocess unstructured text for an unsupervised approach in innovation studies. Moreover, this research provides evidence that our IR model alleviates the semantic ellipsis increasing the chances to find the collaboration concept

2. Experiments

2.1. Data

We use WordNet lexical database (Miller et al., 1990) to select the words referring to the concept of "collaboration" (Table 1). We limit ourselves to the least ambiguous words to reduce the risk of noise that keyword searching would cause. Then, to evaluate our approach we build a test dataset by partially labeling SemCor (Miller & Charles, 1991), a dataset manually annotated with the synsets from WordNet.

Table 1. List of words chosen.

• Consortium • Partnership • Cooperation	• Alliance • Collaborate • Cooperate	• Collaborator • Cooperative • Collaborative

2.2 IR model

An IR model comprises three key components: a query, a target corpus and an IR System. In our case, the query is represented by the list of words referring to collaboration (Table 1), the target corpus is our test dataset and the IR system is based on a cosine similarity computation between the word-embedding vectors representation of the query and the target corpus. Finally, we evaluate the performances of our IR model with the F1-score.

The key elements in our IR approach are the pre-trained word-embeddings, which aim to model the proprieties of a language in a vector space. These techniques leverage neural networks to learn the vector representation of millions of words according to the context in which they appear (Mikolov et al., 2013). The representation of words in dense vector enables the computation of semantically related words and can be used to represent phrases and short texts, reducing the sparsity of traditional vector-space representations (Pelevina et al., 2017). During the years, several pre-trained embedding models were created. We compare the results using GloVe (Pennington et al., 2014) FastText (Bojanowski et al., 2017), and GoogleNews (Mikolov et al., 2013). The three pre-trained models differ in some aspects related to the model architecture and the training corpora used. GoogleNews was the first model to be able to compute high dimensional word vectors from large corpora due to its lower computational complexity. The model is trained on the 100 billion words of the Google news dataset. The backbone is the Skip-gram neural network that predicts the word context giving the word itself (Mikolov et al., 2013). On the other hand, GloVe trained on 840 billion tokens of Common Crawl represents a development on the previous model developed by Mikolov et al. (2013). The model leverages statistical information by training only on the nonzero elements of a word-word co-occurrence matrix (Pennington et al., 2014). Finally, FastText, trained on 600 billion Common Crowl words, improves the GoogleNews algorithm due to its capacity to represent the vector embedding of unseen words as the sum of the vector representations of its n-grams characters.

In the next two sections, we present our results firstly by comparing different settings of the IR model and secondly, by comparing the best IR model with a keyword search method – which is the baseline of this research.

2.3 IR models comparison

To achieve the best performances, we use different parameters for our IR model's three core components generating 24 different settings (with the combination of two queries, three word-embedding models, and four different corpus representations.) For the query, we built the first by transforming each word from Table 1 into a dense vector (we refer to it as W, e.g, GoogleNewsW). From the same list of dense vectors, we then created the second query vector with the arithmetic average (we refer to this one as "Semantic Field" and identify it with the acronym SF e.g., GoogleNewsSF). We justify the last query with the assumption that the SF vector represents the whole semantic field of the concept of "collaboration" which can thus mitigate the problems related to semantic ellipsis. For the target corpus, we execute a preprocessing step of the documents (morphological analysis, lemmatization, removing stop words, etc.), and then we divided each sentence of the target corpus using n-grams, i.e., contiguous word sequences in a document. Indeed, for our target corpus, we tried four different n-grams representations: 1-gram, 2-gram, 3-gram, and 4-gram to find the best setting in our IR model. Finally, for the IR system, we use the three different pre-trained word-embedding models mentioned in section 2.2. Therefore, our IR model works as follows: the IR system takes one word at a time from our list of words as the query, transform it into a dense vector, and searches through the different sentences of the target corpus divided in n-grams, which in turn are transformed into dense vectors, to find the most similar. The similarity is measured by computing the cosine similarity between two dense vectors representing the target corpus and the query. Once the most similar n-gram dense vector is found, the whole sentence is returned as a sentence where the concept of "collaboration" emerges. Finally, since the target corpus is partially annotated, we measure the final results of the IR model by computing the F1-score of the several cosine thresholds to sort out the best ones.

Table 2 shows the F1-score of our several IR models. First, we notice that IR models with W query setting outperform IR models with SF query setting, except for GloVe^{SF} which has a higher F1-score compared to GloVe^{W}. Moreover, we can observe that IR GoogleNews^{w} models get the best performances. In particular, the best model is GoogleNews^{W} with a 3-gram setting, which obtains an F1-score of 0.8635. For the SF query setting, the best model is GoogleNews^{SF} with the 2-gram sequence representation that reaches an F1-score of 0.7926 – which is far less performant than the GoogleNews^{W} model. Additionally, we notice that the performances of the IR models decrease with the 4-gram sequence representation of the target corpus, except for GloVe^{W} which probably requires a longer sequence representation to perform better. It is important to highlight that FastText^{SF} and FastText^{W} perform better in a 1-gram setting, probably due to the subword representation typical to FastText.

Table 2. F1-score of the IR models. The superscript SF indicates that the query setting is the Semantic Field while W indicates the single word; the subscripts indicate the cosine similarity threshold

	1-gram	**2-gram**	**3-gram**	**4-gram**
GoogleNews^{W}	$0.5748_{0.545}$	$0.8540_{0.55}$	$\mathbf{0.8635}_{0.525}$	$0.8385_{0.52}$
Glove^{W}	$0.1835_{0.55}$	$0.1963_{0.55}$	$0.2207_{0.55}$	$0.2375_{0.55}$
FastText^{W}	$\mathbf{0.7532}_{0.55}$	$0.6813_{0.55}$	$0.6048_{0.54}$	$0.4979_{0.53}$
GoogleNews^{SF}	$0.56_{0.51}$	$\mathbf{0.7926}_{0.51}$	$0.7063_{0.50}$	$0.5981_{0.51}$
Glove^{SF}	$0.2870_{0.55}$	$0.3395_{0.55}$	$0.3277_{0.55}$	$0.3040_{0.55}$
FastText^{SF}	$\mathbf{0.6667}_{0.54}$	$0.6064_{0.53}$	$0.45_{0.54}$	$0.3596_{0.54}$

2.4 IR models comparison with baseline

In light of the previous findings, we select the best IR model settings to compare with the method used as the baseline for this study: the keyword search. To ensure that our baseline reaches the best F1-score we apply several pre-processing steps (e.g., lemmatization, lower case, etc.). Additionally, we test combined approaches of keyword search and IR GoogleNews models. Table 3 presents the results of the best models including the measures of Precision and Recall.

Table 3 shows that GoogleNews^{w} with the 3-gram sequence representation outperforms the keyword search in terms of F1-score and Recall. Despite the higher precision of the keyword search, this IR model has a higher F1-score due to its higher number of sentences retrieved (thus, a higher Recall) in which the "collaboration" concept emerges. On the other hand, GoogleNews^{SF} with the 2-gram sequence representation performs worse than the baseline. Among all the methods, the combination keyword search-GoogleNews^{W} with the 4-gram sequence representation yields the highest F1-score in our study. Although this combined method has less Recall than the best GoogleNews^{W} model, it has almost the same precision as the keyword search. Finally, testing the combination of the GoogleNews^{W} with GoogleNews^{SF} models (see Table 3) proved to be the best method to achieve the highest Recall. This result proves that the SF setting, despite its lower performance, significantly contributes to the improved results. In other words, the SF setting and the W setting should be used together since they retrieve different sentences in which the concept emerges.

Table 3. Comparison between IR models, keyword search and combined methods. The superscript SF indicates that the query is the Semantic Field while W indicates that is the single word. The subscripts indicate the cosine similarity threshold

	F1	**Recall**	**Precision**
$GoogleNews^{W}_{0.53}$**4-gram** +$GoogleNews^{SF}_{0.535}$**2-gram**	0.8571	**0.8483**	0.8662
$GoogleNews^{W}_{0.53}$ **4-gram+Keyword**	**0.8736**	0.7862	0.9827
$GoogleNews^{SF}_{0.535}$ **2-gram+ Keyword**	0.8560	**0.7793**	0.9495
Keyword search	**0.8412**	0.7310	0.9965
$GoogleNews^{W}_{0.545}$ **3-gram**	0.8635	0.8069	0.9286
$GoogleNews^{SF}_{0.51}$ **2-gram**	0.7925	0.7379	0.856

3. Conclusion

This research provides a solution to alleviate one of the main issues of the widespread word count methodologies, namely semantic ellipsis. To do so, we compare our baseline, a simple keyword search, with two different methods: an IR model and a combination of IR model with the keyword search. The results show that the IR models mitigate the semantic ellipsis problem outperforming the baseline. However, the combination of the IR models with the keyword search reaches the best performances showing a certain level of complementarity. In particular, the combination of GoogleNews^{w} and GoogleNews^{SF} gets the highest complementarity, thus improving its precision could lead to reach the best performances. Finally, our results suggest that using the methods developed represent an alternative to build stronger and reliable innovation indicators from unstructured text.

Nevertheless, the method developed suffers from two main limitations stemming from the pre-trained models. The first is the impossibility to disambiguate because they conflate all meanings of a word into a single vector (Pelevina et al., 2017). The second limitation is due to the domain sensitivity of the word embedding training corpus that reduce its generalisation.

For our future research directions, we plan to explore four different approaches to improve the results of this study. The first will be to integrate an advanced WSD approach to our combined method. As mentioned above, improving the precision of the IR models will lead to improve their performances. Previous researchers have already used and documented similar combinations of approaches to achieve greater results. For instance, (Rothe & Schütze, 2015) combined word embeddings based on WordNet synsets to obtain sense embeddings, whereas Pina & Johansson (2016) applied random walks on the Swedish Wordnet to generate training data for the Skip-gram model. The second approach will be to use word embedding created with state-of-the-art NLP model as Bert for their capacity to disambiguate words. Indeed, they give to the same word different vector representation based on the context solving one main issue of the pre-trained word-embedding models. The third approach will be to train a supervised model on a manually labeled dataset. This approach united with word embedding could leverage the capacity of the advanced NPL models to disambiguate the words' meanings. Finally, the fourth approach will be to create our word-embedding model based on a collection of documents on innovation studies. We believe that this pre-trained model could have the capacity to learn different concepts and words related to innovation facilitating the creation of new indicators.

References

Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics, 95*(3), 1189–1207.

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PLOS ONE, 16*(4), e0249583. https://doi.org/10.1371/journal.pone.0249583

Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy*, *24*(2), 406–428.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

Héroux-Vaillancourt, M., Beaudry, C., & Rietsch, C. (2020). Using web content analysis to create innovation indicators—What do we really measure? *Quantitative Science Studies*, *1*(4), 1601–1637.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PloS One*, *16*(4), e0249071.

Libaers, D., Hicks, D., & Porter, A. L. (2016). A taxonomy of small firm technology commercialization. *Industrial and Corporate Change*, *25*(3), 371–405.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2017). Making sense of word embeddings. *ArXiv Preprint ArXiv:1708.03390*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pina, L. N., & Johansson, R. (2016). Embedding senses for efficient graph-based word sense disambiguation. *Proceedings of TextGraphs-10: The Workshop on Graph-Based Methods for Natural Language Processing*, 1–5.

Rothe, S., & Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *ArXiv Preprint ArXiv:1507.01127*.