Word Embeddings as First Steps Towards a New Paradigm in HR

Luca Colombo · Intervieweb s.r.l – Zucchetti Group

ABSTRACT

Natural Language Processing (*NLP*) has recently gained much attention for representing and analysing human language computationally. Its applications spread in a pletora of fields from e-mail spam detection to question answering, even to HR Tech and recruiting. But how so? It has various applications and one of them is integrating the more classical searches (filter based) with semantic levels of text comprehension which enables recruiters to have more complete, diversified, unbiased and accurate results. This short article will explain some of the basic ideas behind this new paradigm taking inspiration from one of the simplest models: *Word2Vec* [1-3].

KEYWORDS

AI · NLP · *Word2Vec* · Recruiting · Innovation · HR Tech

INTRODUCTION

The global strive for automation and AI integration in classical (and often repetitive) jobs started gaining traction since three main factors took place: availability of huge amounts of data, mainly collected from the internet, the dramatic increase of computational power and the revamp of mathematical ideas developed in the second half of the last century, which were made actually usable by the new computational capabilities.

Language itself, which was already a deeply studied but a not understood (and still not) subject, began its journey through AI models and ideas. Wikipedia, Google, Twitter (and many others) text-based datasets gave great insights in how individuals communicate concepts, notions and ideas. However, even to this day, language is a very complex matter thus it would be incorrect to say that state of the art AI models[4] fully understand humans. What we might say it's that they greatly help us in tasks that few years back would have taken much more time and effort. Therefore, with careful and extensive research, they enable bias measurement and removal.

These reasons alone make the AI paradigm adoption in HR Tech almost natural and surely very promising. But how does it work? How can we integrate semantics into HR? We will briefly explain *Word2Vec*, one of the basic semantic models. It has be written a lot about *Word2Vec* and much more will be written in the future. Given that reality, as well as a stricly limit pages, there is little hope that I could say much more here that it hasn't said yet. My point here is not to praise *Word2Vec* or bury it, rather to discuss how AI can really bring benefits to the whole HR industry.

WORD2VEC BASICS

Generally the term *Word2Vec* is often used to represent a group of related models which are shallow, two-layer neural networks that are trained to reconstruct linguistic context of words. However we will use it to indicate the one published by *Google* in 2013 for simplicity.

Word2Vec is neither the first[1] nor the last or the best *NLP* model but it had a huge impact on the field. A first fundamental idea is that words can be represented mathematically in a vector space by vectors (*embeddings*), typically of several hundred dimensions and as such manipulated. The *embeddings* are obtained by training the AI model with one key and partially non-trivial assumption: in a corpus, a large and structured set of texts, words found in similar contexts are themselves similar, it means that they share the same *semantic space*. Thanks to

[1] There is considerable prior work, of course. The *Word2Vec* papers cite relatively few papers before 2000, with the exception of Elman[10] (1990) and Harris[6] (1954). The discussion on *Word2Vec* mentions quite a few more on various topics such as distributional semantics (Weaver 1955[7]; Firth 1957[8]), vector spaces (Salton 1975[9]), singular value decomposition (SVD) (Deerwester 1990[10]), embeddings (Pereira 1993[11]), PMI (point wise mutual information) (Church 1990[12]) and similarity estimates (Resnik 1995[13]; Lin 1998[14]).

this idea the model can effectively utilize either of two model architectures to produce a *distributed representation* of words: *continuous bag-of-words* (CBOW) or *continuous skip-gram*. While the first one makes the model able to predict the current word from a window of surrounding context words (under the *bag-of-word assumption* that the order of context words does not influence predictions), the second one does the opposite.

The less distance there is between two vectors (two words) the more they are likely to be related either by similarity or by analogy. The hook analogy is the following: *man* is to *woman* as *king* is to *x*. It is impressive how it can be just used *Word2Vec* and discover that *x* equals *queen*. The model solves analogy tasks like this by trying all words x', in the vocabulary V, and finding the word that maximizes the **equation 1**, where *sim* stands for *similarity* and it is defined in **equation 2** ($\vec{a}$ and $\vec{b}$ are vector representations of, for example, two arbitrary words *a* and *b* while θ is the angle between them).

$$\hat{x} = ARGMAX_{x' \in V}\, sim(x', king + woman - man) \qquad \text{Equation 1.}$$

$$sim(a,b) \equiv cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}|\,|\vec{b}|} \qquad \text{Equation 2.}$$

The intuition for these names comes from expressing the analogy as $\frac{man}{woman} = \frac{king}{queen}$.

INDA HR SEMANTIC PARADIGM

While it is pretty amazing that such a simple method performs this well, what are the advantages and the differences of applying this new paradigm in HR Tech as **INDA** (*INtelligent Data Analysis*) does?

First of all it is different from previous technologies and this is an advantage in itself: it requires a lot more research and development but it can really bring innovation to the whole industry. By analysing applicants' CVs/Resumes as well as job descriptions we are able to have a deeper understanding of the language, thus opening a lot of new possibilities, opportunities and, above all, new features.

While, until now, a recruiter could search candidates through keywords and obviously miss all the applicants whose CV/Resume didn't have those precise keywords in it, with **INDA AI** he would have a much more complete list of pertinent applicants. For example one could just query "*artificial intelligence*" and find all the CVs/Resumes that have the keywords in them AND the ones that have not, but rather share the same *semantic space* (like *machine learning*, *python* and so on), ranked with a pertinence (*semantic*) score.

Therefore it can provide **INDA** users with *content-based recommendation systems*: similar candidates and, potentially, job-candidate matching. These features effectively boost lots of time-consuming tasks and dramatically improve results, helped by the user field knowledge.

However where is the *intelligent* part of **INDA** and its new paradigm? Not only does it effectively analyse complex text-based documents, but it can actually **learn**, making **tailor-made solutions** and **improving over time**.

On the last note i would like to underline the fact that **INDA** was initially developed for *italian* language (*english* language is under development), but it could offer all these features not only for much more languages but also in a cross-language manner with the **multi-language INDA semantic model**. But that's a topic for another time.

REFERENCES

1. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
2. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.
3. Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013.
4. Large collection of NLP models and their performances by category.
5. Elman, Jeffrey L. "Finding structure in time." Cognitive science 14.2 (1990): 179-211.
6. Harris, Zellig S. "Distributional structure." Word 10.2-3 (1954): 146-162.
7. Weaver, Warren, William N. Locke, and A. Donald Booth. "Machine translation of languages." Translation (1955): 15-27.
8. Firth, John R. "A synopsis of linguistic theory, 1930-1955." Studies in linguistic analysis (1957).
9. Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620.
10. Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391-407.
11. Pereira, Fernando, Naftali Tishby, and Lillian Lee. "Distributional clustering of English words." Proceedings of the 31st annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1993.
12. Church, Kenneth Ward, and Patrick Hanks. "Word association norms, mutual information, and lexicography." Computational linguistics 16.1 (1990): 22-29.
13. Resnik, Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." Journal of artificial intelligence research 11 (1999): 95-130.
14. Lin, Dekang. "An information-theoretic definition of similarity." Icml. Vol. 98. No. 1998. 1998.