Open Access

Open Journal of XXX (OJXX)
Volume X, Issue X, 20XX

www.ronpub.com/journals/ojxx
ISSN XXXX-XXXX

Word embeddings for wine recommender systems using vocabularies of experts and consumers

Christophe Cruz [A], Cyril Nguyen Van [B], Laurent Gautier [B]

[A] Laboratoire Le2i, Univ. Bourgogne Franche-Comté, Dijon, France, christophe.cruz@ubfc.fr
[B] Maison des Sciences de l'Homme, Univ. Bourgogne Franche-Comté, 6 Esplanade ERASME, BP 26611, Dijon, France, {cyril.nguyen-van, laurent.gautier}@ubfc.fr

ABSTRACT

This vision paper proposes an approach to use the most advanced word embeddings techniques to bridge the gap between the discourses of experts and non-experts and more specifically the terminologies used by the two communities. Word embeddings makes it possible to find equivalent terms between experts and non-experts, by approach the similarity between words or by revealing hidden semantic relations. Thus, these controlled vocabularies with these new semantic enrichments are exploited in a hybrid recommendation system incorporating content-based ontology and keyword-based ontology to obtain relevant wines recommendations regardless of the level of expertise of the end user. The major aim is to find a non-expert vocabulary from semantic rules to enrich the knowledge of the ontology and improve the indexing of the items (i.e. wine) and the recommendation process.

TYPE OF PAPER AND KEYWORDS

Visionary paper: *recommender system, wine, word embeddings, ontology, expert and non-expert discourses*

1 INTRODUCTION

Readability and understanding of expert discourses by a non-expert consumer are sources of difficulties impacting the act of purchase [18]. This reality is particularly true in the wine sector economy, where organoleptic and sensory properties of wines described by institutions and experts bear a complex semantic burden difficult to apprehend by wine lovers. In this context, the non-expert takes up the expert's speeches without necessarily having a clear understand of them and often adds a hedonistic and evaluative dimension. Thus, he generates a semantic confusion, or he creates his own speech including a terminology built from past experiences but leading to a semantic coverage reduced to a simplistic technical space (e.g. minerality [3][4]) or semantic-reduced specialization with a maximum evaluative space (e.g. drinkable, drinkableness) [5][6]. This paper addresses these issues by bridging the gap between expert and non-expert understanding, and outlines a preliminary method using well-known natural language processing techniques. The wine-specific *AdWine* platform focuses on consumer advice and support. It allows sharing of sensory experience by formalizing the semantics exploiting the terminologies of experts and consumers on qualitative, sensory and emotional dimensions [22][23].

- **Qualitative**: Qualitative assessment refers to conditions in which wines are valued primarily based on characteristics generally associated with the provenance, stylistic characteristics, or varietal characteristics of a wine [2][17].
- **Sensory**: Wine creates sensations in the mouth of astringency, body, burning, balance, tingling, heat and viscosity. All these non-taste sensations are a consequence of oral-tactile stimulation and thus they are as important as the appearance, the aroma and the taste of the wine [9].

- **Emotional**: Silva et al. in [19] studied the relationship between emotional reactions to two similar wines and the sensory properties of each. They suggested that in similar sensory products, the same emotions attract the attention of the consumers. There is no information on how it can be shown that the different sensory attributes of wine induce a different emotional response [12][19].

Although the difficulty of identifying the vocabularies used by both communities (expert and non-expert) should be covered using machine learning, qualifying the semantic relationship between these vocabularies remains difficult to achieve. The *AdWine* platform offers a methodological and technical solution in the form of a recommendation system based on business expertise and consumer practices. Relevant recommendations can be obtained regardless of the level of expertise of the end user. The system aims at extracting and modeling vocabularies used by non-professional consumers from notice pages, forums, etc., by linking these vocabularies to institutional and expert vocabularies. Unfortunately, this process is still done manually and non-exhaustively. The *AdWine* platform is functional, and this paper, as a further contribution, presents our proposition of research into improving automatic data processing with a semantic approach in Word Embeddings area.

The structure of the paper is the following. Section 2 presents the *AdWine* platform. Section 3 details the *AdWine* platform processes. Section 4 provides our propositions and the last section contain the conclusions.

2 THE *ADWINE* PLATFORM

This paper proposes an approach aimed at filling the semantic gap (controlled vocabularies, ontologies, related data) between the descriptive knowledge of the speeches of experts (oenologists, sommeliers, prescribers incarnated by guides and reference journals), the institutional discourses of the professional organizations, and the discourse of neophyte or enlightened amateurs [6]. At the end of this process, these ontologies will be exploited for a high-level wine recommendation system with associated explanations on the selection of characteristics. The following figure shows the operating diagram based on four main processing phases with emphasis on the third phase, the core of the proposal.

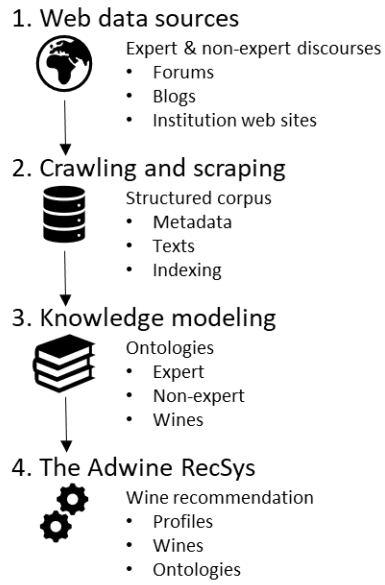


Figure 1 – Sequence of phases starting from data to services

Phase 1 involves selecting relevant data sources from web forums, wine tasting blogs, institutional websites providing wine information, wine tasting guides and journals, and more specifically descriptions of thematic components.

Phase 2 consists of crawling and scraping these data sources to gather texts and to build an expert and non-expert corpus of discourses [25][26]. The objective is to identify in the text wines and their information (name, domain, appellation, producer, vintage, grape variety, type, etc.) as well as the associated discourses of experts and consumers through tasting notices. The contribution of the linguist and the wine expert is required to qualify the representative terms of a specific terminology [6].

Phase 3 aims at modeling the representative terms of a specific terminology as a model of description and classification. It consists of a controlled vocabulary organized as a hierarchy (taxonomy) and detailed in final form of thesaurus. The thesaurus is structured in Resource Description Framework (RDF) format, a graphical model for formally describing resources and their metadata to automatically process such descriptions. An RDF data engine is used to store, query thesauri and have semantic reasoning support. Anything that can be described and conceptualized by the thesaurus, here, a wine, is then considered as an individual. This forms the catalog of items that can be recommended, our knowledge base [27][13].

In a second time this phase which is the proposition of this work consists in processing the texts of the corpus by using Word Embeddings techniques [28] to construct

a single vector space linked to two terminologies (expert and non-expert) helping to find relationships between words such as the example notorious: “king - man + woman ≈ queen”. Finding equivalent terms between experts and non-experts modeled in an ontology enriched with institutional terminologies would provide computational knowledge to recommend more relevant wines with a level of explanation not achieved today.

Phase 4 is the recommendation system based on one of many available approaches. But we propose to use a hybrid approach based on ontologies able to make use of expert and non-expert knowledge. The recommender system computes the distance (or the cosine) between the description vector of an item and a user profile. The proposed method will be based on the qualitative descriptors, sensory and emotional present in both terminological ontologies [24].

Section 3 introduces some background elements on ontology and controlled vocabulary, recommender system, and Word Embeddings. Section 4 discusses the proposition of this paper. We conclude in section 5.

3 FROM DATA TO SERVICES

This section focuses on the representation of controlled vocabularies and the modeling of knowledge from an automatic extraction process of vocabularies using machine learning techniques. In addition, their latent semantics and the recommender systems are discussed. Finally, Word Embeddings techniques are presented.

3.1 Ontology and Controlled Vocabulary

The word “ontology” has different meanings in different communities. However, in computer science, the most popular definition is by Gruber in 1993 [7] who defines Ontology as “an explicit specification of a conceptualization”. This definition is further extended to “Ontologies are a formal, explicit specification of a **shared** conceptualization” by Studer et al., 1998 [20][21]. Thus, ontologies aim at formalizing terms and meanings of knowledge areas. Consequently, ontologies are extremely important in the interaction between systems that constantly exchange information. The proper communication of these systems will only be achieved when both systems receive the same interpretation of the implicit information of the documents exchanged.

Ontology is generally designed to enable the use of a semantic knowledge and application, to facilitate the knowledge sharing process between computers, and to permit the correct semantic interpretation. Formal ontologies using Description Logics [29] are the main component of the Semantic Web. In addition, Semantic Web technologies are based on the open-world hypothesis and take advantage of the inference mechanism. The combination of these two concepts allows the deduction of new knowledge from existing knowledge. Ristoski et al. (2016) [16] describes three use cases where the domains of data mining and the semantic web overlap:

- Use ontologies and Linked Data to support and enhance a knowledge extraction process.
- Use data mining methods to extract knowledge from the semantic web (Semantic web mining).
- Use machine learning methods to generate semantic data and improve Linked Data.

3.2 Recommender Systems

The vast amount of information on the Web, corporate information systems, digital libraries, website sales, and so on, is a well-known fact. Recommender systems aim at providing the best item according to users’ needs. Items can be websites, news articles, books, videos, music, washing machines, etc. In the literature, three paradigms are distinguished, content-based filtering systems based on user networks, collaborative filtering systems based on users’ preferences, and hybrid filtering [1][14]. The third category characterizes Recommender Systems (RS) that combine these first two categories and eventually includes ontologies or any kind of prior knowledge to leverage for instance the cold start issue. For instance, a social network RS should be able to recommend information from its close friends (collaborative filtering) and at the same time the RS should be able to fit the user interest (content-based filtering). Thus, the system can consider a profile as items and provides access to a user’s profile for similar profile search.

Controlled vocabularies are usually used to qualify an item for content-based RS [24]. Therefore, we propose to extend the *AdWine* platform with an expert controlled vocabulary and a non-expert controlled vocabulary. Furthermore, a semantic enrichment process must produce links between both controlled vocabularies. Thus, expert terms related to non-expert terms should be exploitable by hybrid RS incorporating content-based ontology and keyword-based ontology.

How does can work for the wine domain? The recommendation task is mainly based on the comparison or computing the distance between user’s profiles and available item profiles which are in this case the bottles of wine. Thus, classical methods directly used the similarity as a measure of relevance for recommendations. The profile can be seen as the ideal item. Then, more an item is similar to the ideal item

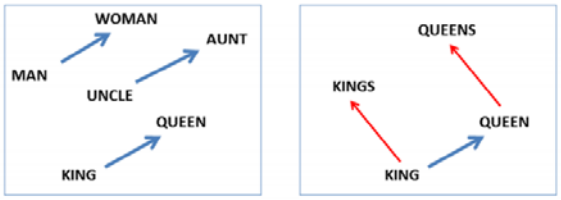
more it is relevant to the user interests. Actually, RS use ontologies and controlled vocabularies are as an index and to provide the item and user profiles. An approach based on the Vector Space Model (VSM) [30] is used in the recommendation task. Wine items and profiles are represented by vectors on a space wherein each dimension is a potential instance of criteria. Several methods can be used to compare vectors; the most common is the cosine similarity. A wine item can be defined as a vector of instances of entities and criteria. A profile can be defined as a vector of instances of criteria [24].

3.3 Machine Learning and Word Embeddings

Since the 1990s, Vector Space Models have been used in distributional semantics with a main drawback which is the co-occurrence matrix sparsity. In parallel, many models for estimating continuous representations of words have been developed, including latent semantic analysis (LSA) and latent Dirichlet allocation (LDA). Contrary to the frequentist approach of VSM, LSA is a Bayesian approach. The most recent and popular methods are named Word Embeddings [10] such as word2vec, a solution that allows for continuous training and the use of pre-trained embeddings, and also GloVe [15], a competitive set of pre-trained word dives. Word embedding is a dense representation of words in the form of vectors. This representation of words allows approximating the similarity between words (i.e., "cat" and "kitten" are similar words, and therefore they are supposed to be close in the reduced vector space) or to disclose hidden semantic relations (i.e., the relationship between "cat" and "kitten" is an analogy to that between "dog" and "puppy").

Contextual information is therefore very useful for learning the meaning and relationship of words because similar words can often appear in the same context. The representation of words is thus done in a space with a form of similarity between them (probabilistic), in which the meaning of the words brings them closer in this space, in terms of statistical distances.

Context creates a space that brings together words that may not have been next to each other in a corpus. These representations make it possible, for example, to find many linguistic regularities simply by performing linear translations in this representation space. For example, the vector result ("Madrid") - vector ("Spain") + vector ("France") gives a position whose nearest vector is vector ("Paris").



(Mikolov et al., NAACL HLT, 2013)

Figure 2 - The lines shown are only mathematical vectors, so we can move "through" by integrating the space of "Man" to "Queen" by subtracting "King" and adding "Woman"

Two main training methods exist:

- « Continuous Bag of Words » (CBOW), resulting in training a neural network to predict a word based on its context, i.e. the words before / after in a sentence.
- « Skip-gram », where we try to predict the context according to the word.

As part of the CBOW, the input of the neural network takes a window around the word and tries to predict the output word. In the context of Skip-gram we try to do the opposite, predict the words around a window determined in advance using the word studied input.

With this vector representation of words, it is possible to use them as features in many basic language processing tasks. It is thus possible to supply conventional algorithms such as a neural network with characteristic vectors of the words.

4 PROPOSITIONS

A focus will be on neural Word Embeddings, the dives of words learned by a neural network, and more particularly the LSTM (Long Short-Term Memory) [8] architecture. There are two well-known algorithms for building "universal" (multi-purpose) embeddings: word2vec from Google (2013) [10][11] and GloVe from Stanford (2014) [15]. Both approaches quantify the similarity between two words by their co-occurrence, their distributional hypothesis. Even though processes use a different method to create the vector space, they produce similar vectors.

word2vec (skip-gram) uses a network of shallow feed-forward neurons (1 hidden layer). For a given input word, the network is trained to predict the probability that each word in the vocabulary will appear next to our input word. In order to accomplish this task, the network has only one weight matrix to adjust that of a dense intermediate hidden layer. Once formed, this hidden layer is the vector representation of words. Word2vec is

a two-layer, shallow neural network, but does not perform deep learning. It can transform plain text into a digital form that deep networks can understand, for example, using recursive neural networks with word connections. GloVe is more traditional because the method uses matrix factorization, and no neural network is involved. It begins through the text and counts the number of times the word pairs are seen near each other in a given window, for instance 10 words. This information is stored in a structure called "co-occurrence matrix". The word vectors are built and adjusted iteratively to minimize the distance (cosine) between the words having a high probability of co-occurrence.

The vectors are very good at answering analogy questions of the form *a* is to *b* as *c* is to*?* For example, *man* is to *woman* as *uncle* is to? (*aunt*) using a simple vector offset method based on cosine distance. It has been found that the similarity of word representations goes beyond simple syntactic regularities. Using a word shift technique where simple algebraic operations are performed on the word vectors, it has been shown, for example, that a vector ("King") - vector ("Man") + vector ("Woman") gives a vector the closest to the vector representation of the word "Queen".

The expert and non-expert discourse corpus is used to create two terminologies related to expert and non-expert terms. Representative terms in the same vector space would be qualified by a linguist and oenologist to distinguish between expert and non-expert terms and classify them in the correct terminology. Based on this fact, the texts of the corpus would be processed to build a single vector space, using Word Embeddings techniques, and only words with the high latent semantics would be examined by experts. Since each term is a vector, it is possible to find equivalent terms between experts and non-experts, to approach the similarity between words or to reveal hidden semantic relations by first using linear translations. Thus, projecting terms in the same vector space open a way to compute expert word vectors close to non-expert word vectors. This mined knowledge about words will enrich the existing terminologies by providing new semantic link between terms.

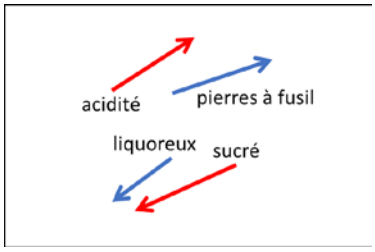


Figure 3 - The lines are the vectors in two dimensions, red arrows are non-expert words and blue arrows are expert ones.

Figure 3 shows how vectors of expert and non-expert vocabularies are closed. We claim this because these vocabularies are often found in same posts discussing about wine tastes.

By defining “semantic rules” to link terms, these relations are modeled in an ontology, using transitions from the vector space to the controlled vocabulary, which helps to search wines in the corpus by using the ontology. The major aspect is to find a non-expert vocabulary from the semantic rules to enrich the knowledge of the ontology and improve the indexing of the items (i.e. wine) and the recommendation process. Another interesting capability is to use element-wise addition of vector elements to ask questions such as ‘Chablis + Chardonnay’ and by looking at the closest tokens to the composite vector come up with impressive answers. Word vectors with such semantic relationships could be used to enhance many existing NLP applications, such as machine translation, information retrieval, and question answering systems.

5 CONCLUSIONS

This visionary paper proposes an approach to use the most advanced Word Embeddings techniques to bridge the gap between the discourses of experts and non-experts and more specifically the terminologies used by the two communities. Highlighting links and modeling this knowledge in an ontology should allow the expert to better understand non-expert consumers. The next phase is to implement a proof of concept in order to demonstrate the feasibility of the approach. First, the expert and non-expert data will be transformed into a vector space using word2vec, then GloVe. This will allow to compare the results of the vector transformation between the two methods. Secondly, from the vector data, and on precise and significant examples of the terms used by the experts, several linear transformations will be tested to discover semantically close terms. The same will be done in the opposite direction based on the terms used by consumers.

This first step opens the door to other services such as:

- Works on the item description generation
- Adapt embeddings to capture the characteristics in a specific domain by using semantic lexicon
- Works on the multi-sense embeddings
- Works on phrases and multi-word expressions
- Works on the temporal dimension of word meanings
- Works on the semantic and the link between qualitative, sensory and emotional dimensions
- Works on the context-words
- Embeddings on multiple languages

ACKNOWLEDGEMENTS

The authors thank the CVT Athéna (http://www.cvt-athena.fr/), the Conseil Régional de Bourgogne-Franche-Comté and the French government for their funding.

REFERENCES

[1] Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. Commun. *ACM 40*(3), 66–72 (Mar 1997).

[2] Danner Lukas, Johnson Trent E., Ristic Renata, Meiselman Herbert L., Bastian Susan E.P., "I like the sound of that!" Wine descriptions influence consumers' expectations, liking, emotions and willingness to pay for Australian white wines, *Food Research International 99* (2017) 263–274.

[3] Deneulin, P., Le Bras, G., Le Fur, Y. & Gautier, L. (2014). Minéralité du vin : représentations mentales de consommateurs suisses et français. *Revue suisse Viticulture Arboriculture Horticulture*, 46(3), pp. 174-180.

[4] Gautier, L., Le Fur, Y., & Robillard, B. (2015). La « minéralité » du vin: mots d'experts et de consommateurs. *In: Gautier, L., & Lavric, E. (Eds). Unité et diversité dans le discours sur le vin en Europe, Frankfurt/Main*, Peter Lang, 149-168.

[5] Gautier, L., & Bach, Matthieu (2017). La terminologie du vin au prisme des corpus oraux de dégustation/présentation (français-allemand) : entre émotions, culture et sensorialité. *Etudes de linguistique appliquée*, 188, pp. 485-509.

[6] Gautier, L. (forthcoming, 2018). La sémantique des termes de dégustation peut-elle être autre chose qu'une sémantique expérientielle et expérimentale ? *In: Verdier, B. & Parizot, A. (Eds). Du Sens à l'Expérience: Gastronomie et Œnologie au prisme de leurs terminologies*, Reims, EPURE.

[7] Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications by A Translation Approach to Portable Ontology Specifications. *Knowledge Creation Diffusion Utilization*, 5:199–220.

[8] Hochreiter S., Schmidhuber J. (1997): Long short-term memory. *Neural Computation*, Volume 9, Issue 8, November 15, 1997, pp. 1735-1780.

[9] Laguna L., Bartolomé B., Moreno-Arribas M. V. (2017), Mouthfeel perception of wine: Oral physiology, components and instrumental characterization, *Trends in Food Science & Technology* 59.

[10] Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations* (ICLR 2013), pp. 1–12.

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, 1–9.

[12] Mora M., Urdaneta E., Chaya C. (2018). Emotional response to wine: Sensory properties, age and gender as drivers of consumers' preferences, *Food Quality and Preference 66*, 19–28

[13] Peixoto, R. Hassan, T., Cruz, C., Bertaux, A. Silva, N. (2015). Semantic HMC: a predictive model using multi-label classification for big data, *Trustcom/BigDataSE/ISPA*, IEEE, 173-179

[14] Portugal I., Alencar P., Donald, C. (2018). The use of machine learning algorithms in recommender systems: *A systematic review*, *Expert Systems with Applications*, Volume 97, 1 May 2018, pp. 205-227.

[15] 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)*, pp. 1532–1543.

[16] Ristoski, P., et Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, 36 :1–22.

[17] Ronald S. Jackson, (2017). Wine Tasting (Third Edition), Chapter 6 – Qualitative Wine Assessment, *A Professional Handbook*, pp. 253–291.

[18] Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers'

online choices. *Journal of retailing*, 80(2), 159-169.

[19] Silva, A. P., van Zyl, H., & Voss, H. (2017). Cheers, proost, saúde: Cultural, contextual and psychological factors of wine and beer consumption in Portugal and in the Netherlands. *Critical Reviews in Food Science and Nutrition*, 57(7), 1340–1349 doi: 10.1080/10408398.2014.969396.

[20] Studer, R., Benjamins, V. R., et Fensel, D. (1998a). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1) :161–197.

[21] Studer, R., Benjamins, V. R., et Fensel, D. (1998b). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197.

[22] Temmerman, R., & Dubois, D. (2017). Food and terminology. Expressing sensory experience in several languages, *Special issue of Terminology*, 23(1).

[23] Valente C., Bauer, F., Venter, F., Watson, B, Nieuwoudt H. (2018). Modelling the sensory space of varietal wines: Mining of large, unstructured text data and visualisation of style patterns, Scientific Reports, 8:4987, doi:10.1038/s41598-018-23347-w.

[24] Werner D., Cruz C. (2013). Precision difference management using a common sub-vector to extend the extended VSM method. *2013 International Conference on Computational Science*. Procedia Computer Science 18, pp. 1179 – 1188.

[25] Hassan T., Cruz C., and Bertaux, A., (2017). Ontology-based approach for unsupervised and adaptive focused crawling. In Proceedings of The International Workshop on Semantic Big Data (SBD '17), Sven Groppe and Le Gruenwald (Eds.). ACM, New York, NY, USA, Article 2, 6 pages. DOI: https://doi.org/10.1145/3066911.3066912

[26] Hassan, T., Cruz, C., Bertaux, A., (2017). Predictive and evolutive cross-referencing for web textual sources, Computing Conference 2017.

[27] Peixoto, R., Cruz, C., Silva, N., (2016). Adaptive learning process for the evolution of ontology-described classification model in big data context, SAI Computing Conference (SAI) 2016, pp. 532-540.

[28] Li Y., Yang T. (2018). Word Embedding for Understanding Natural Language: A Survey. In: Srinivasan S. (eds) Guide to Big Data Applications. Studies in Big Data, vol 26. Springer, Cham

[29] Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., Patel-Schneider, Peter F. (2007). The Description Logic Handbook: Theory, Implementation and Applications, 2nd Edition, Cambridge, ISBN-13: 978-0521150118

[30] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.