Modeling Emergent Associations of Nominal Compounds: Ongoing Research and Preliminary Results*

Massimo Melucci[1] and Laurianne Sitbon[2]

[1] University of Padua, Italy
[2] Queensland University of Technology, Australia

Abstract. This paper presents a roadmap to deliver tools able to model and predict the behaviour of neologisms in the form of nominal compounds. Quite often these compounds yield meanings (in our case, word associations) that are not related to the components of the compounds taken individually. Classical probabilities cannot handle this effect, thus a framework based on quantum probability to model the phenomenon is proposed.

1 Introduction

Most people agree that the combination *pet tree* is likely a *bonzai* and that a *pet human* is probably a *slave*, although none of these combinations exist in the English language, there is no connection between *pet* and the interpretations of the compounds, *bonzai* is a quite rare type of *tree* and *slave* is a rather rare type of *human*. Words that are connected to a combination but are not connected to the single constituent words of the combination are termed "emergent associates". This capacity that humans have to share a creative understanding of novel combinations has been at the centre of several studies in psychology and cognitive science trying to understand what are the processes conducting to their interpretation. In particular, we have evidence that some novel combinations yield to associates that cannot be predicted from the associates of the words taken separately nor from a collection of documents containing both words together.

Information Retrieval (IR) systems can deal with combinations only if they occur in the documents, but they are still largely ignoring novel combinations and emergent associates. In IR, novel combinations not only can be encountered within queries and documents, but also are essential to capture the intentions of authors or users. Multilingual users may express queries using novel combinations assuming a consensus on their interpretation when either they don't know the correct words (paraphrasing) or they translate literally from their native language a concept expressed as a combination. For example, a Chinese user may search

* The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement N. 247590.

documents relevant to *hippopotamus* after literally translating it *river horse*, and *lobster* would be literally translated *dragon shrimp*. This strategy could work in human to human communication, but automatic systems are to date not able to infer such new meanings as early work on cross-lingual information retrieval using dictionary based translation has proven [1]

Additionally, native speakers of a language may voluntarily create novel combinations that cannot be classically interpreted (e.g. from a combination of the associates of the words taken individually) in order to emphasize the importance of the emerging associates and attract the attention of the reader, in the same way neologisms are intended for [2]. News headlines are typical examples of application of such strategies, with headlines such as *Oklahoma Surprise: Islam as an Election Issue* or *Brother's transplant gift carries unbearable cost*, as well as video titles with for example *fish wish* or a Wallmart adventure that attracted a lot of attention on a single day. More generally, it was also found "that no less than 39% of the neologisms had idiomatic meanings at their very birth" [3].

The previous examples of usage of novel combinations reflect the main problems (among others) QONTEXT (EU funded contract N. 247590) is striving to address: (*i*) can we predict whether a combination will yield to an emergent associate? (*ii*) can we predict what will be the meaning of an novel combination?

2 Research Question and Contribution

The key to solve these problems is to find a representation that allows for the modelling of the emergence of unexpected associates from novel combinations. Classical representations of words and their meaning are either based on semantic networks (formal relations) or semantic spaces (based on co-occurrences). Composition is then expressed as a function of the neighbours or the dimensions of each word taken individually within a single space [4] which doesn't allow for the modelling of emergent associates. QONTEXT will explore the suitability of quantum probability to model the combined representations as well as its ability to act as a predictive model.

Quantum Theory (QT) has been previously shown to be the only way of going beyond the classical interpretation of compositionality to solve the problem of typicality of combinations [5]. This problem is classically known as the guppy effect because a guppy is found not to be a typical example of pet nor fish, but highly typical of "pet fish". The goal is not only to model existing combination and examples, but also to derive a computational model that can be used to interpret novel combinations enough to generate query expansion terms or accurate translations. To this end, a quantum probability space has to be defined. In this paper, it is shown that a single quantum probability space in one vector space can be defined.

3 Words, Vectors and Probability

In the classical probabilistic model, events (e.g., associates or combinations) are represented as sets and the probability measure is based on a set measure, e.g., set cardinality. In contrast, in quantum probability, events are represented as subspaces and probabilities are generated by density matrices[3], the most simple case being vectors. Suppose that the vectors $|w\rangle$, $|b\rangle$ represent an event and a density, respectively.[4] The probability that w occurs given b is provided by the *quantum probability rule*, that is, $|\langle w|b\rangle|^2 = |a_{wb}|^2$ where a_{wb} is termed amplitude and $|\langle x|x\rangle|^2 = 1$ for every $|x\rangle$.[5]. In general, the probability measure is given by the trace of the product between the density matrix which corresponds to the probability distribution, and the matrix representing an event. A quantum probability space is then given by a set of subspaces and a density matrix. [6]

4 Combination and Quantum Probability

Suppose that $b = b_1 b_2$ is a combination (e.g. *pet tree*) and v is an associate (e.g., *slave* or *bonzai*). Using classical probability, these events are sets and then $b \subseteq b_1, b_2$ (if b occurs, then both b_1 and b_2 occur but the viceversa does not hold). Suppose also that b_1 and b_2 are quite common words, thus their probability is not negligible. Moreover,

$$P(v, b_1) \geq P(v, b_1, b_2) \geq P(v, b) \qquad P(v, b_2) \geq P(v, b_1, b_2) \geq P(v, b)$$

If v (e.g., *bonzai*) rarely occurs with b_1 (e.g., *pet*) or with b_2 (e.g., *tree*), it is also rare within b and thus it cannot be an emergent associate, which cannot be detected by classical probability.
Using quantum probability, instead,

$$|b\rangle = a_{ib}|b_i\rangle + \bar{a}_{ib}|\bar{b}_i\rangle \qquad |a_{ib}|^2 + |\bar{a}_{ib}|^2 = 1 \qquad \langle b_i|\bar{b}_i\rangle = 0 \qquad i = 1, 2$$

and

$$\begin{aligned} |\langle v|b\rangle|^2 &= |a_{ib}\langle v|b_i\rangle + \bar{a}_i b\langle v|\bar{b}_i\rangle|^2 \\ &= |a_{ib}|^2|a_{iv}|^2 + |\bar{a}_{ib}|^2|\bar{a}_{iv}|^2 + 2I \qquad i = 1, 2 \end{aligned} \tag{1}$$

where the sum of the first two terms of the right-hand side is the classical probability $P(v|b)$ and

$$I = |a_{ib}||a_{iv}||\bar{a}_{ib}||\bar{a}_{iv}| \cos\theta \tag{2}$$

is the interference term. The conditions for emergence can be studied using (1) and (2). As $-1 \leq I \leq +1$, (1) can be lower or higher than $P(v|b)$ and then can be used for predicting emergent associates. The estimation of I is thus crucial. In particular, whereas the the $|a_{ij}|^2$'s are

[3] In general, density operators.
[4] The Dirac notation is used in QT.
[5] $\langle x|$ is the transpose of $|x\rangle$.

empirical probabilities, the estimation of $\cos\theta$ is the most difficult step because it is mainly due to the interpretation of θ, which is the angle of the complex number $a_{ib}a_{iv}\bar{a}_{ib}\bar{a}_{iv}$ – further investigation are then needed. The vector formalism allows us to build the quantum probability space. Suppose that the empirical probabilities $|a_{ij}|^2$ are estimated for n associates $w_1, \ldots, w_n$ with respect to the b_i's, $i = 1, 2$. If the $|a_{ij}|^2$'s are arranged in a $2 \times n$ matrix,

$$(|w_1\rangle \cdots |w_n\rangle) = \left(|b_i\rangle \ |\bar{b}_i\rangle\right) \begin{pmatrix} a_{i1} & \cdots & a_{in} \\ \bar{a}_{i1} & \cdots & \bar{a}_{in} \end{pmatrix} \qquad i = 1, 2 \tag{3}$$

Through SVD of the $2 \times n$ matrix, the $|w_j\rangle$'s and the $|b_i\rangle$'s can be built, thus obtaining the space and the probability $|\langle b_1|b_2\rangle|^2$ which measures the distance between the constituent words of the combination b.

5 Concluding Remarks and Future Work

A single representation of the events is obtained so that the empirical probabilities can be reproduced by the quantum probability rule. The modeling results are still quite preliminary, however, we are confident that the ongoing research may provide the useful theoretical framework for detecting emergent associates.

References

1. Kwok, K.: Evaluation of an english-chinese cross-lingual retrieval experiment. In: Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval. (1997) 110–114
2. Lehrer, A.: Understanding trendy neologisms. Italian Journal of Linguistics (2003)
3. Schmid, H.J.: New words in the mind: Concept-formation and entrenchment of neologisms. Anglia-Zeitschrift Fur Englische Philologie **126** (2008) 1–36
4. Lapata, M., Mitchell, J.: Composition in distributional models of semantics. Cognitive Science **34** (2010) 1388–1429
5. Aerts, D., Gabora, L.: Diederik aerts and liane m. gabora, a theory of concepts and their combinations ii: A hilbert space representation. Kibernetes **34** (2004) 192–221
6. Melucci, M., van Rijsbergen, C.J.: Quantum mechanics and information retrieval: a probabilistic overview. In: Advanced Topics in Information Retrieval. Springer (Forthcoming)