A Quantum-Like Approach to Word Sense Disambiguation

Fabio Tamburini
FICLIT - University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

This paper presents a novel algorithm for Word Sense Disambiguation (WSD) based on Quantum Probability Theory. The Quantum WSD algorithm requires concepts representations as vectors in the complex domain and thus we have developed a technique for computing complex word and sentence embeddings based on the Paragraph Vectors algorithm. Despite the proposed method is quite simple and that it does not require long training phases, when it is evaluated on a standardized benchmark for this task it exhibits state-of-the-art (SOTA) performances.

1 Introduction

The introduction of the *prototype theory* by E. Rosch (1973), one of the most influential theories describing concept organisation at cognitive level, completely changed the perspective in semantics and nowadays most of the studies in computational semantics consider concepts membership and concepts similarity as graded features in a "semantic space".

In Natural Language Processing (NLP) all the recent and fundamental studies on word and sentence embeddings, e.g. (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016; Le and Mikolov, 2014; Sutskever et al., 2014; Kiros et al., 2015; Subramanian et al., 2018; Cer et al., 2018), as well as the older works on word spaces based on co-occurence measures, e.g. see the reviews from (Turney and Pantel, 2010; Baroni and Lenci, 2010), rely on an high-dimensional vector space to represent concepts, through the application of the Harrisian distributional hypothesis, collecting contextual information from text corpora, and measuring their relationships by means of some kind of geometric distance.

The fundamental experiments of Tversky (1977) in cognitive psychology on concept similarity and concept combination challenged the common view to adopt the Classical, Kolmorogovian, Probability Theory (CPT) to explain and model these phenomena. Starting from Tversky's data, a large set of newer experimental studies in cognitive psychology showed a systematic violation of the basic axioms of CPT. For example, the classical problem known as the "pet-fish problem" or "guppy-effect" (Osherson and Smith, 1981) is a typical case of overextension in concept conjunction: if we denote as *pet-fish* the conjunction of the concepts *pet* and *fish* and ask people to rate "guppy" as member of *pet*, *fish* and *pet-fish*, they tend to consider it as a highly typical *pet-fish* but neither a particularly typical *pet* nor *fish*. This violates the CPT axiom stating that the probability of the events conjunction must be less or equal to the probability of the single events. In a similar way scholars in cognitive psychology proposed various experiments showing a systematic violations of CPT axioms in concept disjunction, conceptual negation, decision making and some other relevant cognitive processes: see, for example, (Hampton, 1988; Alxatib and Pelletier, 2011; Aerts et al., 2015). These studies show clearly the impossibility of modeling such cognitive phenomena by using CPT and even by using further elaboration of it such as the fuzzy-set probability theory.

In the same field a growing set of studies in the last decades started to explore the possibility of modeling such cognitive phenomena by using different, more sophisticated, probability theories, in particular Quantum Probability Theory (QPT), the foundational calculus of Quantum Mechanics Theory (QMT). We refer the reader to these books or comprehensive reviews for an in-depth introduction to these approaches in cog-

Proceedings of Recent Advances in Natural Language Processing, pages 1176–1185,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_135

nitive psychology (Busemeyer, 2012; Haven and Khrennikov, 2013; Pothos and Busemeyer, 2013; Yearsley et al., 2015; Wendt, 2015; Ashtiani and Azgomi, 2015; Aerts et al., 2016a,b; Haven and Khrennikov, 2018). A very large set of these works showed that, by applying the axioms of QPT and taking advantage from the peculiar phenomena modeled by this calculus, such as *superposition*, *interference* and *entanglement*, it is possible to build complete models which are able to well explain all the experimental data and incorporate in the theory all the deviations from CPT exhibited by human behaviours during cognitive experiments. Even if some studies try to find connections by QPT and brain functionality at neural level (Khrennikov et al., 2018), the use of QPT in this field is simply as an explanation theory useful to model real phenomena in the right way, but none of them is really claiming that our brain is working by applying QPT axioms. That is why it is common to use the term "quantum-like" to describe models making use of this calculus in cognitive psychology.

Even if QMT is one of the most successful theories in modern science, the attempts to apply it in other domains remain rather limited, excluding, of course, the large quantity of studies regarding Quantum Information Processing on quantum computers and Electronics. Only in recent years some scholars tried to embody principles derived from QMT into their specific fields, for example, by the Information Retrieval community (van Rijsbergen, 2004; Zuccon et al., 2009; Melucci and van Rijsbergen, 2011; González and Caicedo, 2011; Melucci, 2015), by the Economics and Finance community (Baaquie, 2018) and by the community studying Quantum Computation and Information Theory (Nielsen and Chuang, 2010; Wilde, 2013). In the machine learning field (Arjovsky et al., 2016; Wisdom et al., 2016; Jing et al., 2017) have used unitary evolution matrices to build deep neural networks obtaining interesting results, but we can observe that their works do not completely adhere to QPT and use unitary evolution operators in a way not allowed by QPT. (Moreira and Wichert, 2018) presented an interesting application of QPT for developing Quantum Bayesian Networks. In recent years, also the NLP community started to look at QPT with interest and some studies using it have already been presented (Blacoe et al., 2013; Liu et al., 2013; Kartsaklis et al., 2016; Basile and Tamburini, 2017).

Given this general framework, and the successful results in explaining cognitive experiments, it seemed natural trying to explore the possibility of applying QPT also in NLP and see if the peculiar properties of this calculus could introduce some benefits for solving NLP tasks.

In this paper we will apply QPT for modeling the Word Sense Disambiguation problem testing our proposal on well-known benchmarks. WSD is an historical task which aims to assign the correct word sense for a polysemous word given a linguistic context. The possible senses for a given word are extracted from a reference sense inventory. We refer the reader to the reviews of (Agirre and Edmonds, 2007; Navigli, 2009; Vidhu Bhala and Abirami, 2014) and to the papers describing the last evaluation results (Navigli et al., 2013; Moro and Navigli, 2015; Raganato et al., 2017a) to get a clear picture of the SOTA for this task.

Computational systems solving such task can be broadly divided into two main groups: knowledge-based systems do not require a sense-annotated corpus for training the model and are usually based on lexical or knowledge resources for performing the disambiguation process; on the contrary, supervised WSD systems require a sense annotated corpus in order to train the model and set up all the parameters. Looking at the previously cited evaluations, supervised WSD systems are able to produce the best results and are currently establishing the SOTA: see, for example, (Zhong and Ng, 2010; Iacobacci et al., 2016; Papandrea et al., 2017; Raganato et al., 2017b; Tripodi and Pelillo, 2017; Luo et al., 2018b,a; Melacci et al., 2018; Uslu et al., 2018). Despite these long time studies, the Most-Frequent-Sense baseline is still a strong algorithm challenging all new proposals, and the best systems results are only few points over that baseline.

2 Quantum Probability Theory

This section aims to introduce the basic background knowledge necessary to understand QPT and the underlying mathematical constructions. A more complete introduction about these topics can be found, for example, in (Nielsen and Chuang, 2010; Busemeyer, 2012). It is important to note that QPT is a probability theory more general than CPT and it includes it completely.

Quantum Events

QPT assigns probability to events as well as classical Kolmogorovian probability theory, but, unlike CPT that defines events as sets, it defines events as subspaces of a multidimensional complex Hilbert space $\mathcal{H} = \mathbb{C}^n$.

Quantum States

In QPT the state of a quantum system is defined, using the Dirac notation[1], as a complex vector $|\psi\rangle \in \mathcal{H}$ with $||\psi\rangle| = 1$ and, in its general formulation, it can be expressed as

$$|\psi\rangle = \phi_1 |e_1\rangle + \phi_2 |e_2\rangle + ... + \phi_n |e_n\rangle \quad (1)$$

where ϕ_j are complex numbers called *probability amplitudes*, $\phi_j = \langle e_j|\psi\rangle$, and $\{|e_j\rangle\}$ is a basis of the Hilbert space $\mathcal{H}$. The state in (1) is called a *superposition* state w.r.t. the basis vectors seen as basic states.

For each event subspace spanned by the vector $|x\rangle$, it is possible to build a projector operator $P_x = |x\rangle\langle x|$ that can project a generic state vector $|\psi\rangle$ onto the subspace corresponding to that event.

Quantum Measurements

In QPT, quantum measurements of a variable (or observable) M are usually represented by a set of measurement operators $\{M_k\}$ where the index indicates one of the possible measurement outcomes and $\sum_k M_k^\dagger M_k = I$ (I denotes the $n \times n$ identity matrix). Applying a measurement on a quantum system when in state $|\psi\rangle$ we can compute the probability of getting a specific result k as

$$P(k) = \langle\psi|M_k^\dagger M_k|\psi\rangle\,. \quad (2)$$

When we measure a quantum system and an event is observed, the act of measuring it changes the state of the system from the superposed state $|\psi\rangle$ to a new state, it is said that the system *collapses*; this new state is given by

$$|\psi\rangle' = \frac{M_k\,|\psi\rangle}{\sqrt{\langle\psi|M_k^\dagger M_k|\psi\rangle}}. \quad (3)$$

An important class of measurements is known as projective measurements. These measurements are represented by Hermitian observables that admit a spectral decomposition $M = \sum_k v_k P_k$ where $P_k = |u_k\rangle\langle u_k|$ is the projector onto the eigenvector $|u_k\rangle$ of M with eigenvalue v_k and we can compute the probability of obtaining the result k as $P(k) = \langle\psi|P_k|\psi\rangle = |\langle u_k|\psi\rangle|^2$. The eigensystem obtained by the spectral decomposition imply the assumption of orthogonality between the eigenvectors and thus force measurements on this orthonormal basis of $\mathcal{H}$. Once applied a measurement the system will collapse and no more uncertainty will remain. A further measurement of the outcome k will result in $P(k) = 1$.

There is also another type of measurements, the Positive Operator-Valued Measurement (POVM). Projective measurements require the assumption of orthogonality and are not well suited to measure non-orthonormal states and compute their probabilities. A POVM is a set of Hermitian positive operators $\{E_i\}$ such that $\sum_i E_i = \sum_i M_i^\dagger M_i = I$ is the only requirement. Note that for projective nonorthogonal operators all M_i can be written as outer products of general, non orthonormal, state vectors and we can introduce any number of operators E_i.

Quantum Interference

Interference is one of the most intriguing phenomena arising only in the domain of quantum systems. The classical double slit experiment is often used as a simple example to introduce this phenomenon, see, for example, (Zuccon et al., 2009). Let us shoot a physical particle towards a screen with two slits A and B and, once passed the screen, the particle hits a detector panel behind the screen in a specific position x. By closing one of the two slits, say B, we can compute the probability that the particle hits the detector at a position x passing through A, $P_A(x) = |\phi_A(x)|^2$, or the reverse, by closing A, $P_B(x) = |\phi_B(x)|^2$ where $\phi_A(x)$ and $\phi_B(x)$ are the probability amplitudes associated with the two events $|e_A\rangle$ and $|e_B\rangle$ forming an orthonormal basis for $\mathcal{H} = \mathbb{C}^2$. By applying the classical probability we can compute $P_{AB}(x)$ when both slits are open and the particle can pass either through A or B as

$$P_{AB}(x) = P_A(x) + P_B(x) = |\phi_A(x)|^2 + |\phi_B(x)|^2 \quad (4)$$

but experimentally we can note that this equality does not hold and that we have to correct equation (4) by applying the QPT adding an *interference*

[1] In Dirac notation $|.\rangle$ is a column vector, or a *ket*, while $\langle.|$ is a row vector, or a *bra*. Using this notation the inner product between two vectors can be expressed as $\langle x|y\rangle$ and the outer product as $|x\rangle\langle y|$. Then $\langle x| = |x\rangle^\dagger$, where † marks the conjugate transpose operation on vectors or matrices.

term:

$$\begin{aligned} P_{AB}(x) &= |\phi_A(x)|^2 + |\phi_B(x)|^2 + I_{AB}(x) \\ &= |\phi_A(x)|^2 + |\phi_B(x)|^2 + \\ &\quad (\phi_A(x)^* \phi_B(x) + \phi_A(x)\phi_B(x)^*) \end{aligned}$$

In summary, the classical Kolmogorovian rule for addition of probabilities when the event can occur in various alternative ways is violated and we have to apply the QPT in order to completely explain the experiments results.

3 Quantum WSD

3.1 Background

The literature on cognitive psychology gives us precious suggestions about the definitions of the elements involved in our problem and on how operationalise them in the framework of QPT.

Concepts can be seen as quantum states described by state vectors, $|\psi\rangle$, in a complex Hilbert space $\mathcal{H}$.

Specific entities or exemplars or, more appropriately in the WSD domain, polysemous words are viewed as superposed states between the referring senses, or concepts, state vectors. For example the vector for a polysemous word $|W\rangle$ can be expressed as

$$|W\rangle = \phi_1 |S_1\rangle + ... + \phi_m |S_m\rangle \qquad (5)$$

where $\{|S_j\rangle\}$ represents the set of all its possible sense vectors.

A context, as a piece of text in a natural language (e.g. a sentence), provides a specific meaning to a polysemous word collapsing its superposition to one of its possible senses. It is described as a measurement operation projecting the system state into a specific subspace spanned by the linguistic context.

In order to transform these general intuitions into a practical system, the crucial step regards the possibility of generating vector representations of words and senses in the complex domain. The large set of works introducing word and sentence embeddings (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016; Le and Mikolov, 2014; Sutskever et al., 2014; Kiros et al., 2015; Subramanian et al., 2018; Cer et al., 2018) produce representations in the real domain while we need similar vectors but in the complex domain. The next section will show how to transform a classical word embedding approach in order to obtain complex word/sentence embeddings.

3.2 Complex Word/Sentence Embeddings

There are some recent work in literature (Trouillon et al., 2016; Li et al., 2018) proposing techniques for computing complex-valued sentence embeddings to solve a specific task by training a Deep Neural Network (DNN) on the problem output classes. Although strictly connected with our work, these studies learn complex embeddings tailored to a specific task, while we are looking for a procedure to learn general complex representations of words and sentences potentially useful for solving a wide range of tasks.

We took inspiration from the unpublished underdocumented code made available by Théo Trouillon[2] extending `word2vec` code[3] (Mikolov et al., 2013) for working on complex numbers and generating complex word embeddings.

In `word2vec` the skip-gram negative-sampling model is trained minimising the objective

$$E = -log\ \sigma(\mathbf{v'}_{w_O}^T \mathbf{v}_{w_I}) - \sum_{w_j \in \mathcal{W}_{neg}} log\ \sigma(-\mathbf{v'}_{w_j}^T \mathbf{v}_{w_I})$$

where σ is the sigmoid function, w_O is the output word (a positive sample taken from the real context), $\mathbf{v'}_{w_O}$ is its corresponding vector taken from the output weight matrix and $\mathbf{v}_{w_I}$ is the value of the hidden layer that, for skip-gram models, is equivalent to the input word (w_I) vector taken from the input weight matrix.

For extending this model to work with complex numbers we have to transform the input and output weight matrices from real to complex values and adapt the objective function consequently. Unfortunately there are no studies, up to our knowledge, handling directly complex losses and most of the more recent attempts to work with complex neural networks (Trabelsi et al., 2018; Scardapane et al., 2018; Sarroff, 2018) transform the complex loss into a real one by applying some function f to the network output. We can then adapt the objective function as following

$$\begin{aligned} E' &= -log\ \sigma\Big(f\left(\langle \mathbf{v'}_{w_O} | \mathbf{v}_{w_I} \rangle\right)\Big) - \\ &\quad \sum_{w_j \in \mathcal{W}_{neg}} log\ \sigma\left(f\big(-\langle \mathbf{v'}_{w_j} | \mathbf{v}_{w_I} \rangle\big)\right) \end{aligned}$$

where $\mathbf{v'}_{w_x}$ and $\mathbf{v}_{w_x}$ are now complex vectors and

[2] https://github.com/ttrouill/imwords.git

[3] https://code.google.com/archive/p/\word2vec/

$f : \mathbb{C} \to \mathbb{R}$ can be defined as

$$f(z) = \Re(z) + \Im(z) = \frac{z + \overline{z}}{2} + \frac{z - \overline{z}}{2i}.$$

where $\overline{z}$ is the complex conjugate of z.

E' is still a real-valued loss function and, by leveraging the Wirtinger calculus to extend the complex derivative (well defined only for holomorphic functions) to non-holomorphic functions and real-valued analytic functions, we can calculate and backpropagate all the gradients needed to update the network weights:

$$\frac{\partial E'}{\partial \mathbf{v}'_{w_j}} = (1 + i) \left[\sigma \left(\langle \mathbf{v}'_{w_j} | \mathbf{v}_{w_I} \rangle \right) - t_j \right] \mathbf{v}_{w_I}$$

$$\frac{\partial E'}{\partial \mathbf{v}_{w_I}} = (1 - i) \left[\sigma \left(\langle \mathbf{v}'_{w_j} | \mathbf{v}_{w_I} \rangle \right) - t_j \right] \mathbf{v}'_{w_j}$$

where $t_j = 1$ if $w_j = w_O$ (positive sample), $t_j = 0$ if $w_j \in \mathcal{W}_{neg}$ (negative samples) and i is the imaginary unit.

Le and Mikolov (2014) proposed an extension to the `word2vec` model to build vectors referred to a generic piece of text (phrases, sentences, paragraphs or entire texts). They called this extension "Paragraph Vectors" (PVs). Following their paper and the suggestions given by T. Mikolov for implementing PVs[4], we extended the code accordingly, producing complex paragraph vectors (cPVs) for fragments of texts longer than a word. As in the cited paper, it was sufficient to insert a fake word at the beginning of the paragraph and training it together with all the other words forming the paragraph to obtain, at the end of the training process, reliable dense vector representations for the paragraph in the complex domain as well as complex vector representations for words (cWV). Although this vectors are not derived using QPT and we used the Dirac notation only for consistency, they will enable us to use such complex vectors as the basic elements in our WSD algorithm based on QPT.

3.3 The WSD Model

The proposed model for WSD relies heavily on an external lexical resource for getting all the glosses and examples connected to a specific meaning. WordNet (Miller, 1995), BabelNet (Navigli and Ponzetto, 2012) or other lexical resources providing a large set of senses with their glosses and examples can be used for our purpose.

Given the general considerations made in Sections 3.1 and the complex vector representations introduced in 3.2, we can list the ingredients for our WSD recipe in the following way:

- the target word W to be disambiguated will be represented as the corresponding cWV, namely $|W\rangle$;
- the subspace of $\mathcal{H}$ connected with the sense S, has to be build by combining all the glosses and examples provided by the external lexical resource, seen as the corresponding cPVs, and all the disambiguated contexts extracted from the training set belonging to this specific sense, again seen as cPVs. The whole set of vectors $\{|\mathcal{G}_j\rangle\}$ belonging to a specific sense S are, in general, non-orthogonal each other and thus cannot form a proper basis to define the subspace connected with the sense. A standard procedure to obtain an orthonormal basis spanning the same subspace of a specific set of vectors is based on the Singular Value Decomposition and it is available in any linear algebra library. Given the orthonormal basis spanning the same space of $\{|\mathcal{G}_j\rangle\}$, say $\{|\mathcal{O}_i\rangle\}$, we can build the projector over the subspace spanned by $\{|\mathcal{G}_j\rangle\}$ relative to the sense S as

$$P_S = \sum_i |\mathcal{O}_i\rangle\langle\mathcal{O}_i| \tag{6}$$

- the context subspace will be represented by all the cPVs corresponding to the sentences belonging to the context and the projector P_C to this subspace can be computed following the same procedure as in the previous point.

Having defined such ingredients, the disambiguation process consists in projecting the word state $|W\rangle$ onto the context subspace by applying the quantum measurement operation of (3)

$$|W_C\rangle = \frac{P_C |W\rangle}{\sqrt{\langle W | P_C^\dagger P_C | W \rangle}}$$

and then compute, by applying another measurement on the new state $|W_C\rangle$, which of the possible

[4] https://groups.google.com/d/msg/word2vec\-toolkit/Q49FIrNOQRo/J6KG8mUj45sJ

senses of W, $\{S_k\}$, exhibits the maximum similarity with $|W_C\rangle$ or, in other words, the projection of $|W_C\rangle$ over S_k has the maximum probability:

$$\overline{S} = \underset{S_k}{\operatorname{argmax}}\, P(S_k) = \underset{S_k}{\operatorname{argmax}}\, \langle W_C | P_{S_k}^{\dagger} P_{S_k} | W_C \rangle$$

where P_{S_k} is the projector obtained by equation (6) for sense S_k.

4 Experiments

We made two kind of experiments: the first is aimed to evaluate if the proposed procedure to learn complex word/sentence embeddings from texts produces effective results, while the second is devoted to the specific evaluation of our Quantum WSD system (QWSD). Both experiments rely on standard, largely-used evaluations benchmarks.

4.1 Complex Embedding Evaluation

Producing complex sentence embeddings for getting the best performance is not the main focus of this work. We simply need word/sentence representations in the complex domain in order to use QPT to develop our new approach to WSD. Thus, the evaluation of the cPVs is simply devoted to be certain that the cVPs are reliable dense representations of our glosses and contexts sentences.

To test the cPVs we adopted the benchmark proposed by (Conneau and Kiela, 2018) to evaluate sentence embeddings focusing on the five Semantic Textual Similarity (STS) tasks. We chose to apply only these tasks because we are not interested in a complete evaluation, but only in getting a broad idea if our cPVs were reliable enough to build our QWSD model.

Table 1 shows the evaluation results. The performances of our model in the STS tasks are in line with the other basic method for producing sentence embeddings. For a fair comparison, cPVs have the same number of parameters as the other methods, thus, considering that the experiments in (Conneau and Kiela, 2018) were made with vectors of 300 dimensions, our result is referred to complex embeddings with 150 dimensions.

4.2 QWSD Evaluation

In order to evaluate the proposed method to solve the WSD problem we relied, as most of the recent studies, on the standardized evaluation proposed by (Raganato et al., 2017a) for English all-words WSD. This benchmark is based on two

Model	STS				
	'12	'13	'14	'15	'16
GloVe BoW	0.52	0.50	0.55	0.56	0.51
fastText BoW	0.58	0.58	0.65	0.68	64.3
SkipThought-LN	0.31	0.25	0.31	0.31	-
InferSent	0.59	0.59	0.70	0.71	0.72
Char-phrase	0.66	0.57	0.75	0.76	-
ELMo (Orig.5.5B)*	0.55	0.53	0.63	0.68	0.60
USE (DAN)*	0.59	0.59	0.68	0.72	0.70
USE (Transf.)*	0.61	0.64	0.71	0.74	0.74
PVs (300)	0.53	0.61	0.66	0.69	0.65
cPVs (150)	0.53	0.61	0.65	0.69	0.64

Table 1: Evaluation of sentence representations on the STS benchmarks as the average of Pearson correlations. Systems marked with * use embeddings bigger than 300 dim. Data were taken from (Conneau and Kiela, 2018) and (Perone et al., 2018). At the end the results of the cPVs and the standard PVs in the real domain.

corpora for training the systems, namely SemCor (Miller et al., 1994) and OMSTI (Taghipour and Ng, 2015) and five test corpora taken from Senseval/SemEval evaluation campaigns: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013) and SemEval-2015 (Moro and Navigli, 2015).

We compare our evaluation results with all the systems already evaluated by (Raganato et al., 2017a) and with the new studies presented in the last two years (Papandrea et al., 2017; Zhong and Ng, 2010; Iacobacci et al., 2016; Raganato et al., 2017b; Luo et al., 2018b,a; Melacci et al., 2018; Uslu et al., 2018).

Most of the studies cited before required complex training phases and they use the SemEval-2007 dataset as the validation set, thus, although we did not need to use it in this way, it has to be excluded from the test sets for evaluating WSD systems. Moreover, most of the previous results were obtained by using only SemCor as training set and we stick to this practice to enable a complete comparability of the various results. The standard metric for this task is the F-score.

The setting for our experiments is very simple:

- we collected the glosses and the examples for a given sense, a Wordnet synset, from BabelNet v3.7.
- with regard to the creation of cPVs, by following the unsupervised procedure described in 3.2, we created a single corpus formed by all the sentences contained in the British

National Corpus[5] joined with the BabelNet glosses and examples for the various senses and all the training and test sets contexts to be used during the evaluation. It is important to underline that we connected the training set context with the correct target word sense, but for the test set we simply connected the contexts to their instance id without any explicit link to the correct results. In other words, the fake words we inserted for generating the cPVs are the correct WordNet sense id string for the training context and the test instance id string for the test contexts. In this way we can retrieve the cPVs when needed without compromising the evaluation. We used the hyperparameters setting proposed in the Mikolov's post cited before without any parameter optimisation (emb. size = 400, win. size = 10, neg. samples = 5, sub-sampl. = 1e-4, iter. = 20, min. word freq. = 5).

- the disambiguation procedure is deterministic and does not have any parameter to tune. We only introduced a limit to the number of senses for each target word equal to 20.

4.3 Results

Table 2 shows the results obtained by the proposed system (QWSD) compared with the results obtained by the SOTA systems on the evaluation framework proposed by (Raganato et al., 2017a). QWSD, despite its simplicity, obtained very good results, not far from those obtained by the best systems on the same benchmark, and it exhibits the best performances in the last evaluation datasets, namely SemEval 2013 and 2015, the best performance when classifying polysemous nouns and the second best for adjectives.

5 Discussion and Conclusions

After the influential paper from Reimers and Gurevych (2017) it is clear that we should report the mean and standard deviation of various runs with the same setting in order to get a more accurate picture of the real systems performances. We had no possibility to reproduce all the results from the other systems because some of them lack of a public code, others do not work well or it is not sufficiently clear how to set up them and for others the results we obtained using the public code and the described parameters are so different from the published ones that, to be fair with the colleagues, we prefer not to take any position on it and thus we put in Table 2 our best result, as some other studies did. In any case, to be consistent with the community trend, we made ten different experiments to generate the cPVs and repeated the training procedure accordingly. Our results over the ten runs are very similar to the best one: 70.1 ± 0.22. The problem of results reproducibility for empirical studies is becoming rather serious (Wieling et al., 2018).

But, why our method is working well? A possible explanation for these results could be traced back to the interference phenomenon. A superposition of state vectors is usually a valid state vector even if they are not orthogonal, thus we can consider the vector representing a polysemous word, $|W\rangle$ as in (5), as a superposition state. As showed by (Khrennikov and Basieva, 2014; Aliakbarzadeh and Kitto, 2016) this can still produce the interference phenomenon even if these vectors are non-orthogonal. The presence of the interference term when computing the word sense probability over the context subspace might explain the good results we obtained. This point deserves further studies in order to verify this idea and understand how to use this term to drive the disambiguation process and obtain even better results.

Our Quantum WSD system relies on complex vector representations for words and sentences; in this study, for ease of experimentation, we tested our proposal by using a simple extension of PVs to the complex domain, but in literature there are techniques to build better word/sentence embeddings that could be extended to the complex domain; the field of complex DNN is very active.

Another interesting idea worth to be explored regards the possibility of solving the disambiguation process for all ambiguous words in the sentence as a single process (Tripodi and Pelillo, 2017) by using specific properties of QMT.

We feel it is worth spending few words on the simplicity of the proposed system. The only training phase regards the production of cPVs and, as well as the standard `word2vec` application, is based on a very simple feedforward neural network employing very few non-linearities. The disambiguation phase based on QPT is fully deterministic and involves few linear algebra operations, namely matrix multiplications and orthogonalisation procedures. Looking at the perfor-

[5] http://purl.ox.ac.uk/ota/2554

System	Test datasets				All Test datasets				
	SE2	**SE3**	**SE13**	**SE15**	**Noun**	**Verb**	**Adj**	**Adv**	**ALL**
Most Frequent Sense baseline	65.6	66.0	63.8	67.1	67.6	49.6	73.1	80.5	65.5
IMS (Zhong and Ng, 2010)	70.9	69.3	65.3	69.5	70.4	56.1	75.6	82.9	68.9
IMS+emb (Iacobacci et al., 2016)	71.0	69.3	67.3	71.3	71.8	55.4	76.1	82.7	69.7
IMS-s+emb	72.2	70.4	65.9	71.5	71.9	56.9	75.9	84.7	70.1
Bi-LSTM+att+lex (Raganato et al., 2017b)	72.0	69.4	66.4	72.4	71.6	57.1	75.6	83.2	69.9
Bi-LSTM+att+lex+pos	72.0	69.1	66.9	71.5	71.5	57.5	75.0	83.8	69.9
supWSD (Papandrea et al., 2017)	71.3	68.8	65.8	70.0	-	-	-	-	69.1
supWSD+emb	72.7	70.6	66.8	71.8	-	-	-	-	70.6
supWSD-s+emb	72.2	70.3	66.1	71.6	-	-	-	-	70.1
GAS (Linear) (Luo et al., 2018b)	72.0	70.0	66.7	71.6	71.7	57.4	76.5	83.5	70.1
GAS (Conc)	72.1	70.2	67.0	71.8	72.1	57.2	76.0	84.4	70.3
GAS_ext (Linear)	72.4	70.1	67.1	72.1	71.9	58.1	76.4	84.7	70.4
GAS_ext (Conc)	72.2	70.5	67.2	72.6	72.2	57.7	76.6	**85.0**	70.6
CAN^w (Luo et al., 2018a)	72.3	69.8	65.5	71.1	71.1	57.3	76.5	84.7	69.8
CAN^s	72.2	70.2	69.1	72.2	73.5	56.5	76.6	80.3	70.9
HCAN	72.8	70.3	68.5	72.8	72.7	58.2	**77.4**	84.1	71.1
fastSense (Uslu et al., 2018)	73.5	**73.5**	66.2	73.2	-	-	-	-	71.7
IMSC2V+PR (Melacci et al., 2018)	73.8	71.9	68.2	72.8	73.1	77.1	60.6	83.5	71.8
IMSC2V+sSyn	**74.2**	71.8	68.1	72.8	71.9	76.2	57.6	83.2	**71.9**
IMSC2V+sSyn+PR	74.1	71.6	68.1	72.8	73.1	**77.3**	60.2	83.8	71.8
QWSD	70.5	69.8	**69.8**	**73.4**	**73.6**	54.4	77.0	80.6	70.6

Table 2: Results obtained by the proposed system (QWSD) compared with the SOTA (F-score). The first four columns show the results for the different test sets, while the last five the performances on all the four test sets joined together analysed w.r.t. the different parts of speech.

mances in Table 2 it is clear that the results are very near, and in some case better, than those obtained by system based on intricate DNN structures that require long training processes and a careful parameter tuning. This paper presents the results of a basic quantum system for WSD and the results are very encouraging; more work in this direction could drive to even better systems.

Codes and data are freely available[6].

References

D. Aerts, J. Broekaert, L. Gabora, and S. Sozzo. 2016a. Generalizing prototype theory: A formal quantum framework. *Frontiers in psychology* 7:418.

D. Aerts, J. Broekaert, L. Gabora, and S. (Eds.) Sozzo. 2016b. *Quantum Structures in Cognitive and Social Science*. Frontiers Media, Lausanne.

D. Aerts, S. Sozzo, and T. Veloz. 2015. Quantum structure of negation and conjunction in human thought. *Frontiers in Psychology* 6:1447.

E. Agirre and P. Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company.

M. Aliakbarzadeh and K. Kitto. 2016. Applying povm to model non-orthogonality in quantum cognition. *Lecture Notes in Computer Science* 9535:284–293.

S. Alxatib and J. Pelletier. 2011. On the psychology of truth-gaps. In R. Nouwen, R. van Rooij, U. Sauerland, and H. Schmitz, editors, *Vagueness in Communication*. Springer, Berlin Heidelberg, pages 13–36.

M. Arjovsky, A. Shah, and Y. Bengio. 2016. Unitary evolution recurrent neural networks. In *Proc. of the ICML 2016*. pages 1120–1128.

M. Ashtiani and M. Abdollahi Azgomi. 2015. A survey of quantum-like approaches to decision making and cognition. *Mathematical Social Sciences* 75:49–80.

B.E. Baaquie. 2018. *Quantum Field Theory for Economics and Finance*. Cambridge University Press.

M. Baroni and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36:673–721.

I. Basile and F. Tamburini. 2017. Towards quantum language models. In *Proc. of EMNLP 2017*. pages 1840–1849.

W. Blacoe, E. Kashefi, and M. Lapata. 2013. A quantum-theoretic approach to distributional semantics. In *Proc. of HLT-NAACL 2013*. pages 847–857.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

J.R. Busemeyer. 2012. Introduction to quantum probability for social and behavioral scientists. In L. Rudolph and J. Valsiner, editors, *Qualitative Mathematics For the Social Sciences*, London: Routledge, pages 75–104.

[6] https://github.com/ftamburin/QWSD

D. Cer, Y. Yang, S. Kong, et al. 2018. Universal sentence encoder for english. In *Proc. of EMNLP 2018*. pages 169–174.

A. Conneau and D. Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proc. of LREC 2018*.

P. Edmonds and S. Cotton. 2001. Senseval-2: Overview. In *Proc. of SENSEVAL-2*. pages 1–5.

F.A. González and J.C. Caicedo. 2011. Quantum latent semantic analysis. In *Advances in Information Retrieval Theory, LNCS, 6931*. pages 52–63.

J.A. Hampton. 1988. Disjunction of natural concepts. *Memory & Cognition* 16(6):579–591.

E. Haven and A. Khrennikov. 2013. *Quantum Social Science*. Cambridge University Press, Cambridge.

E. Haven and A. (Eds.) Khrennikov. 2018. *Applications of Quantum Mechanical Techniques to Areas Outside of Quantum Mechanics*. Frontiers Media.

I. Iacobacci, M.T. Pilehvar, and R. Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proc. of ACL 2016*. pages 897–907.

L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S.A. Skirlo, M. Tegmark, and M. Soljacic. 2017. Tunable efficient unitary neural networks (EUNN) and their application to RNN. In *Proc. of ICML 2017*.

D. Kartsaklis, M. Lewis, and L. Rimell. 2016. *Proc. of the 2016 Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science*.

A. Khrennikov and I. Basieva. 2014. Quantum model for psychological measurements: From the projection postulate to interference of mental observables represented as positive operator valued measures. *NeuroQuantology* 12(3):324–336.

A. Khrennikov, I. Basieva, E.M. Pothos, and I. Yamato. 2018. Quantum probability in decision making from quantum information representation of neuronal states. *Scientific Reports* 8(1):16225.

R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. 2015. Skip-thought vectors. In C. Cortes et al., editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 3294–3302.

Q.V. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of ICML'14*. pages 1188–1196.

Q. Li, S. Uprety, B. Wang, and D. Song. 2018. Quantum-inspired complex word embedding. In *Proc. of The Third Workshop on Representation Learning for NLP*. pages 50–57.

D. Liu, X. Yang, and M. Jiang. 2013. A novel classifier based on quantum computation. In *Proc. of ACL 2013*. pages 484–488.

F. Luo, T. Liu, Z. He, Q. Xia, Z. Sui, and B. Chang. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proc. of EMNLP 2018*. pages 1402–1411.

F. Luo, T. Liu, Q. Xia, B. Chang, and Z. Sui. 2018b. Incorporating glosses into neural word sense disambiguation. In *Proc. ACL 2018*. pages 2473–2482.

S. Melacci, A. Globo, and L. Rigutini. 2018. Enhancing modern supervised word sense disambiguation models by semantic lexical resources. In *Proc. of LREC 2018*.

M. Melucci. 2015. *Introduction to Information Retrieval and Quantum Mechanics*. The IR Series 35. Springer, Berlin Heidelberg.

M. Melucci and K. van Rijsbergen. 2011. Quantum mechanics and information retrieval. In M. Melucci and R. Baeza-Yates, editors, *Advanced Topics in Information Retrieval*, Springer, Berlin Heidelberg, pages 125–155.

T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges et al., editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

G.A. Miller. 1995. Wordnet: A lexical database for english. *Communiations of the ACM* 38:39–41.

G.A. Miller, M. Chodorow, S. Landes, C. Leacock, and R.G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proc. of HLT'94*. pages 240–243.

C. Moreira and A. Wichert. 2018. Are quantum-like bayesian networks more powerful than classical bayesian networks? *Journal of Mathematical Psychology* 82:73–83.

A. Moro and R. Navigli. 2015. Semeval-2015 task 13: Multilingual all words sense disambiguation and entity linking. In *Proc. of SemEval'15*. pages 288–297.

R. Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.* 41(2):1–69.

R. Navigli, D. Jurgens, and D. Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proc. of SemEval'13*. pages 222–231.

R. Navigli and S.P. Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

M.A. Nielsen and I.L. Chuang. 2010. *Quantum Computation and Quantum Information*. Cambridge University Press, New York.

D.N. Osherson and E.E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9(1):35 – 58.

S. Papandrea, A. Raganato, and C. Delli Bovi. 2017. Supwsd: A flexible toolkit for supervised word sense disambiguation. In *Proc. of EMNLP 2017*. pages 103–108.

J. Pennington, R. Socher, and C.D Manning. 2014. Glove: Global vectors for word representation. In *Proc. EMNLP 2014)*. pages 1532–1543.

C.S. Perone, R. Silveira, and T.S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR* abs/1806.06259.

E.M. Pothos and J.R. Busemeyer. 2013. Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences* 36(3):255–274.

S. Pradhan, E. Loper, D. Dligach, and M. Palmer. 2007. Semeval-2007 task-17: English lexical sample, SRL and All words. In *Proc. of SemEval'07*. pages 87–92.

A. Raganato, J. Camacho-Collados, and R. Navigli. 2017a. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proc. of EACL 2017*. pages 99–110.

A. Raganato, C. Delli Bovi, and R. Navigli. 2017b. Neural sequence learning models for word sense disambiguation. In *Proc. of EMNLP 2017*. pages 1156–1167.

N. Reimers and I. Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proc. of EMNLP 2017*. pages 338–348.

E. Rosch. 1973. Natural categories. *Cognitive Psychology* 4:328–350.

A.M. Sarroff. 2018. *Complex Neural Networks for Audio*. Ph.D. thesis, Dartmouth College, Hanover, NH.

S. Scardapane, S. Van Vaerenbergh, A. Hussain, and A. Uncini. 2018. Complex-valued neural networks with nonparametric activation functions. *IEEE Transactions on Emerging Topics in Computational Intelligence*. pages 1–11.

B. Snyder and M. Palmer. 2004. The english all-words task. In *Proc. of SENSEVAL-3*. pages 41–43.

S. Subramanian, A. Trischler, Y. Bengio, and C.J. Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *Proc. of ICLR 2018*.

I. Sutskever, O. Vinyals, and Q.V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS'14*. pages 3104–3112.

K. Taghipour and H.T. Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proc. of CoNLL 2015*. pages 338–344.

C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. Felipe Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C.J. Pal. 2018. Deep complex networks. In *Proc. of ICLR 2018*.

R. Tripodi and M. Pelillo. 2017. A game-theoretic approach to word sense disambiguation. *Computational Linguistics* 43(1):31–70.

T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. 2016. Complex embeddings for simple link prediction. In *ICML 2016*. pages 2071–2080.

P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188.

A. Tversky. 1977. Features of similarity. *Psychological Review* 84(4):327–352.

T. Uslu, A. Mehler, D. Baumartz, A. Henlein, and W. Hemati. 2018. FastSense: An efficient word sense disambiguation classifier. In *Proc. of LREC 2018*.

K. van Rijsbergen. 2004. *The Geometry of Information Retrieval*. Cambridge University Press, New York.

R. V. Vidhu Bhala and S. Abirami. 2014. Trends in word sense disambiguation. *Artificial Intelligence Review* 42(2):159–171.

A. Wendt. 2015. *Quantum Mind and Social Science: Unifying Physical and Social Ontology*. Cambridge University Press.

M. Wieling, J. Rawee, and G. van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics* 44(4):641–649.

M.M. Wilde. 2013. *Quantum Information Theory*. Cambridge University Press.

S. Wisdom, T. Powers, J. Hershey, J. Le Roux, and L. Atlas. 2016. Full-capacity unitary recurrent neural networks. In D.D. Lee et al., editors, *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., pages 4880–4888.

J.M. Yearsley, E.M. Pothos, J.A. Hampton, and A.B. Duran. 2015. Towards a quantum probability theory of similarity judgments. *LNCS* 8951:132–145.

Z. Zhong and H.T. Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proc. of ACL 2010 Demos*. pages 78–83.

G. Zuccon, L.A. Azzopardi, and K. van Rijsbergen. 2009. The quantum probability ranking principle for information retrieval. In L.A. Azzopardi et al., editors, *Advances in Information Retrieval Theory*, Springer, Berlin Heidelberg, pages 232–240.