Quantum-inspired Representation for Long-tail Senses of Word Sense Disambiguation

Anonymous submission

Abstract

Data imbalance, also known as the long-tail distribution of data, is an important challenge for data-driven models. In the Word Sense Disambiguation (WSD) task, the long-tail phenomenon of word sense distribution is more significant, making it difficult to effectively represent and identify Long-Tail Senses (LTSs). Therefore exploring representation methods that do not rely heavily on the training sample size is an important way to combat LTSs. Considering that many new states, that is, superposition states, can be constructed from several known states in quantum mechanics, superposition states provide the possibility to obtain more accurate representations from inferior representations learned from a small sample size. Inspired by quantum superposition states, a representation method in Hilbert space is proposed to reduce the dependence on large sample sizes and thus combat LTSs. We theoretically prove the correctness of the method, and verify its effectiveness under the standard WSD evaluation framework and obtain state-of-the-art performance. Furthermore, we also test on the constructed LTS and latest cross-lingual datasets, and achieve promising results.

1 Introduction

Data imbalance is very common in the field of data mining (Abd Elrahman and Abraham 2013). For real data, the sample sizes of different categories are generally not uniform distributions, but unbalanced, such as the most common long-tailed distribution. Large datasets often exhibit such a long tail phenomenon (Yang and Xu 2020). For data-driven models, data imbalance directly leads to difficulty in obtaining effective representation and recognition for tail categories (Ntoutsi, Fafalios, and Gadiraju 2020).

Word Sense Disambiguation (WSD) is to determine the word sense of the target word according to the given context, which belongs to the basic research topic in natural language processing (Bevilacqua et al. 2021; Navigli 2009). But for high-level tasks based on natural language understanding, an accurate identification of word senses at the lexical level can effectively improve the overall performance. Due to the long tail phenomenon of word sense distribution in linguistics, the Long-Tailed Senses (LTSs) lack sufficient training samples to complete effective feature representation, and then obtain correct recognition (Aliwy and Taher 2019; Bevilacqua et al. 2021).

Researchers employ data augmentation methods, such as oversampling (Fujii et al. 1998), data synthesis (Bolshina and Loukachevitch 2020), multilingualism (Scarlini, Pasini, and Navigli 2020a), etc., to improve the disadvantaged position of LTSs. These methods are not universally effective and destroy the statistical laws inherent in the data. Methods of changing learning strategies have also been extensively studied, such as transfer learning (Kohli 2021), meta-learning (Holla et al. 2020), few-shot learning (Kumar et al. 2019), etc.

Kumar et al. (Kumar et al. 2019) adjusted the discrete label space often used in the WSD task into a continuous sense embedding space, resulting in state-of-the-art performance in experiments. The success of this scheme is not only because the sense embeddings replace the original labels, but also more importantly because the **continuous space constraint** helps to obtain correct sense representations. Compared with the Euclidean space without constraints, continuous space constraints can compress the range of parameter values and reduce the number of parameters.

Considering that quantum states are constructed in Hilbert space, it naturally has the advantage of continuous space constraints. In addition, considering that many new quantum states, that is, **superposition states** (Nielsen and Chuang 2002), can be constructed from several known quantum states in quantum mechanics, the concept of superposition states provides the possibility to obtain more accurate representations from inferior representations learned from small sample sizes. In this article, we leverage quantum states to describe representations learned from insufficient training samples, and further obtain more accurate representations in their quantum state form by constructing superposition states.

In quantum mechanics, quantum states, including superposition states, are points on the sphere of Hilbert space (the special case of entangled states is not discussed here), so the representations described by quantum states have the advantage of continuous space constraints. Furthermore, learning the representations does not require adding additional parameters. Because building a superposition state requires an additional parameter, the continuous space constraint can save a parameter. In general, the quantum-inspired representation method provides the ability to obtain more accurate representations under continuous space constraints without

increasing the amount of parameters. See Sec. 3.2 and 4.2 for theoretical proofs and specific implementation details respectively.

Our contributions can be summarized as follows:

- Propose a method to obtain more accurate representations in the case of insufficient training samples, inspired by quantum superposition states, which can alleviate the dependence of data-driven models on large sample sizes;
- Theoretically prove that the representation method is better than the classical one;
- Implement experiments under the standard evaluation framework for the WSD task and obtain state-of-the-art performance; Execute experiments under the constructed long-tail sense and latest cross-lingual datasets and also achieve promising results.

2 Related Work

2.1 Long-tail Word Sense Disambiguation

With the continuous optimization and improvement of intelligent algorithms, high-frequency word senses can be accurately identified. At this stage, the focus of the WSD community is gradually converging on long-tail word sense disambiguation (Bevilacqua et al. 2021; Aliwy and Taher 2019).

Blevins et al. (Blevins and Zettlemoyer 2020) first realized that the focus of the current WSD task should be on LTSs, and proposed joint training of dual encoders to enhance the representation of word sense with the knowledge of training samples. Kumar et al. (Kumar et al. 2019) paid attention to the representation and recognition of zero-shot word sense disambiguation, and proposed to use continuous word sense embedding space to replace discrete label space to deal with unseen word senses. In addition, there are Refs. (Wang and Wang 2021; Wang, Zhang, and Wang 2021a; Berend 2020).

The representation method proposed in this article is specially designed for the scarcity of training samples of LTSs, and also adopts the characteristics of continuous space, namely Hilbert space.

2.2 Quantum-inspired Models for WSD

Quantum theory is believed to be able to reveal human cognitive behavior (Busemeyer and Bruza 2012; Bruza, Wang, and Busemeyer 2015), and the mathematical principles of quantum mechanics (namely, quantum probability theory) have been extensively studied because of their superiority (Li et al. 2018; Liu, Hou, and Song 2021).

On the WSD task, Tamburini (Tamburini 2019) adopted quantum probability theory to calculate the similarity between the complex word or sentence embeddings. Although the model is relatively simple, it has the advantage that it does not take a lot of time during the training phase. In addition, Kumar et al. (Kumar et al. 2019) adjusted the discrete label space into a continuous sense embedding space, resulting in state-of-the-art performance in the experiments. Although this work does not mention quantum probability theory, the pure state in quantum mechanics is also established in a continuous space, indicating that continuous space constraints are indeed beneficial to the representation and recognition of word senses.

Inspired by the above work, this article leverages the mathematical form of quantum superposition state to obtain an optimal representation under the premise of suboptimal representation, that is, to search for a better representation on the basis of the original representation.

3 Theoretical Analysis

3.1 Preliminaries

This section introduces the background knowledge of Quantum Probability Theory (QPT) necessary to understand theoretical proof and model design. *Note that QPT is a more general probability theory and is perfectly compatible with Classical Probability Theory (CPT), so it can be used as a modeling tool in information systems. See Ref. (Nielsen and Chuang 2002) for more details.*

Quantum Events: QPT assigns probabilities to events like CPT, but it defines events in the subspace of the multi-dimensional complex Hilbert space $\mathcal{H} \in \mathbb{C}^n$, unlike CPT that defines events as sets.

Quantum States: In QPT, the quantum system (i.e., quantum state) is defined as a complex vector $|\psi\rangle \in \mathcal{H}$ with $\||\psi\rangle\| = 1$ using the Dirac[1] notation. A more general formalization can be defined as

$$|\psi\rangle = \phi_1|e_1\rangle + \phi_2|e_2\rangle + ... + \phi_i|e_i\rangle + ... \tag{1}$$

where $(\phi_1)^2 + (\phi_2)^2 + ... + (\phi_i)^2 + ... = 1$, ϕ_i is a complex number called the probability amplitude, $\phi_i = \langle e_i|\psi\rangle$, and $|e_i\rangle$ is the basis of the space $\mathcal{H}$. This state is called a **superposition state**, and its basis vectors are the basic states. It is also possible to construct a superposition state from other superposition states,

$$|\psi\rangle = \phi_1|\psi_1\rangle + \phi_2|\psi_2\rangle + ... + \phi_i|\psi_i\rangle + ... \tag{2}$$

Quantum Measurements: There is more than one type of quantum measurement in quantum mechanics, such as general measurement, projection measurement and POVM measurement (Nielsen and Chuang 2002). The textbook mainly introduces projection measurement, while the field of quantum information processing mostly uses general measurement.

General measurement is described by a set of measurement operators $\{M_m\}$, where m refers to the possible result. For example, the status of the system is $|\psi\rangle$, and the probability of the measured result m is

$$p(m) = p(m; M_m) = \langle\psi|M_m^\dagger M_m|\psi\rangle. \tag{3}$$

After the measurement, the status of the system changes to

$$|\psi'\rangle = \frac{M_m|\psi\rangle}{\sqrt{\langle\psi|M_m^\dagger M_m|\psi\rangle}}. \tag{4}$$

[1] In Dirac notation, $|\cdot\rangle$ is a column vector (or called *ket*), while $\langle\cdot|$ is a row vector (or called *bra*). Using these symbols, the inner product can be expressed as $\langle x|y\rangle$ and the outer product as $|x\rangle\langle y|$. Also $\langle x| = |x\rangle^\dagger$, where "†" marks the conjugate transpose operation on vectors or matrices.

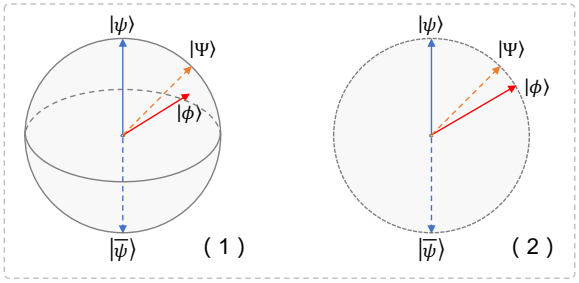

Figure 1: Illustration of the superposition state on the Bloch sphere. (1) refers to high-dimensional space and (2) to plane.

Note that $\{M_m\}$ needs to satisfy the completeness, $\sum_m M_m^\dagger M_m = I$, which reveals the fact that the sum of the probabilities is 1,

$$\sum_m p(m) = \sum_m \langle\psi|M_m^\dagger M_m|\psi\rangle = 1. \tag{5}$$

3.2 Theoretical Analysis for Quantum Inspired Representation

When the word or text vector ψ obtained with an insufficient training sample size is represented by the quantum state $|\psi\rangle$, we can obtain a more correct representation through the superposition state constructed by $|\psi\rangle$ and its anti-state $|\overline{\psi}\rangle$ (the method of obtaining $|\overline{\psi}\rangle$ is given in Sec. 4.2). The formal definition and proof of this conclusion are as follows.

Conclusion 1. *Assuming that $|\phi\rangle$ is the correct state and $|\psi\rangle$ is obtained when the training sample size is insufficient, the superposition state*

$$|\Psi\rangle = \cos(\theta)|\psi\rangle + \sin(\theta)|\overline{\psi}\rangle \tag{6}$$

can be better than $|\psi\rangle$, that is, there is

$$p(|\psi\rangle; |\phi\rangle\langle\phi|) \leq p(|\Psi\rangle; |\phi\rangle\langle\phi|) \leq 1 \tag{7}$$

where $|\phi\rangle\langle\phi|$ represents a measurement operator, $M_\phi = |\phi\rangle\langle\phi|$, and this probability operation can be defined as

$$p(|\cdot\rangle; M_\phi) = \langle\cdot|M_\phi^\dagger M_\phi|\cdot\rangle = \langle\cdot|\phi\rangle\langle\phi|\cdot\rangle = (\langle\phi|\cdot\rangle)^2 \tag{8}$$

by general measurement theory.

Proof. In the high-dimensional Hilbert space, the superposition states composed of $|\psi\rangle$ and its anti-state $|\overline{\psi}\rangle$ can form a plane, which is a great circle under the Bloch sphere description method.

- When $|\phi\rangle$ is **not on the plane** formed by the superposition states, as shown in Fig. 1 (1), some superposition states on the plane are closer to $|\phi\rangle$ than $|\psi\rangle$, i.e., $p(|\psi\rangle; M_\phi) \leq p(|\Psi\rangle; M_\phi)$.
- When $|\phi\rangle$ is **on the plane** formed by superposition states, as shown in Fig. 1 (2), $|\phi\rangle$ can be characterized by one of the superposition states, i.e., $p(|\Psi\rangle; M_\phi) = 1$, and in other cases, it is $p(|\Psi\rangle; M_\phi) < 1$.

□

Note that the proof shows that some better representations can be obtained based on superposition states, but it does not deny that there are worse cases. However, in specific applications, the algorithm can finally obtain a more correct representation described by the superposition state through optimization learning.

*Moreover, it needs to be emphasized that this article uses the method of increasing the system dimension to replace the **complex number representation** of the quantum system, and their functions are equivalent in the form of a single quantum system (Hardy 2001; Janotta and Hinrichsen 2014).*

4 Methodology

This section is divided into four parts: a formal description of the WSD task is given in Sec. 4.1; the quantum-inspired representation method is formalized in Sec. 4.2; the WSD model under the Quantum-inspired Representation (called **QR-WSD**) is constructed in Sec. 4.3; the loss function and optimization method are given in Sec. 4.4.

4.1 Word Sense Disambiguation

WSD belongs to a standard classification task, and its core task is to learn a mapping model from the target word to word senses. When using pre-trained language models to vectorize the target word and word senses, the core task is to match the most similar word sense embeddings emb^i_{sense} for the target word embedding emb_{target},

$$Min\{..., d(emb_{target}, emb^i_{sense}), ...\} \tag{9}$$

where $d(\cdot)$ represents a distance metric function.

4.2 Quantum-inspired Representation

In some tasks of Natural Language Processing (NLP), texts or words are vectorized, such as text or word embedding $V = [..., v_i, ...]$, and on this basis, we can construct quantum systems (i.e., quantum states) by imposing the constraint

$$\mathcal{C}(V) = |V\rangle = \frac{1}{\sqrt{\sum_i v_i^2}} V \tag{10}$$

with $\||V\rangle\| = 1$. By the constraint $\mathcal{C}$, all vectors can be represented as quantum systems. The system generated by this method does not conflict with the system constructed from the basis vectors, that is, it can be decomposed into a composite form of the basis vectors, as shown in Eq. (1). Note that quantum states of single systems, i.e., points on the sphere of Hilbert space, can be obtained by this method (Nielsen and Chuang 2002).

Based on the above quantum system, the superposition state needed in this article can be constructed,

$$|\mathbf{V}\rangle = \cos(\theta)|V\rangle + \sin(\theta)|\overline{V}\rangle \tag{11}$$

where $\theta \in \mathbb{R}$ and $|\overline{V}\rangle = \mathbf{X}|V\rangle$. $\mathbf{X}$ is the NOT gate of quantum mechanics, and its high-dimensional form is shown as

$$\mathbf{X} = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}. \tag{12}$$

For the parameter θ in Eq. (11), we can obtain it in many ways in practice, such as setting it as an independent parameter, or setting it as a non-independent parameter learned based on other features. We here obtain θ from V through a linear layer of the neural network,

$$\mathcal{L}(V) = \theta \in \mathbb{R}. \tag{13}$$

Therefore, the mapping from word or text vectors to superposition states can be formalized as

$$\mathcal{F}(V) = \cos(\mathcal{L}(V))\mathcal{C}(V) + \sin(\mathcal{L}(V))\mathbf{X}\mathcal{C}(V) = |\mathbf{V}\rangle. \tag{14}$$

At this point, we can use quantum measurement theory to calculate the similarity between the representation obtained from the superposition state and the representation of the label. Since the output of the quantum measurement is the probability, it can be directly used as the score of the corresponding label. Here we define the correct state $|\phi\rangle$ (which can be the vectorized form of the label in specific applications) as a measurement operator $M_\phi = |\phi\rangle\langle\phi|$, and the probability of $|\mathbf{V}\rangle$ can be obtained from the general measurement theory,

$$p(|\mathbf{V}\rangle; M_\phi) = \langle \mathbf{V}|M_\phi^\dagger M_\phi|\mathbf{V}\rangle = (\langle\phi|\mathbf{V}\rangle)^2. \tag{15}$$

4.3 Architecture of QR-WSD

Based on the quantum-inspired representation method, a WSD model is constructed, called QR-WSD, and its architecture is shown in Fig. 2.

The overall structure of the model can be divided into two parts, namely **the classical processing method part** and **the quantum processing method part**. Previous work (Kumar et al. 2019; Blevins and Zettlemoyer 2020) has verified that the classic method can effectively deal with head senses (that is, most frequent senses). This paper will integrate the quantum method, which is good at processing tail senses, to make up for the shortcomings of previous work. Both quantum and classical processing methods rely on the embeddings extracted by BERT (Devlin et al. 2019) from the target word, glosses and example sentences, where glosses and example sentences come from WordNet (Miller 1998). *It should be emphasized that although this article uses the glosses and example sentences, the difference is that we only use them for training label vectors instead of directly using them as training data.* In the end, these two methods together determine the final output of the model.

Classical Processing Method The classical processing method part uses the embeddings obtained by the two BERTs as input. One BERT is used to extract the target word and example sentence embeddings (*in Fig. 2, for the convenience of understanding, we repeat the BERT*), and the other BERT to extract gloss embeddings. The generation process of the embeddings is described below, and they are also applicable to the quantum processing method part.

- **Target word embedding**: The training set is used as the input of BERT to obtain the embedding of the target word. The text is processed following the unified regulations of BERT, such as adding $[CLS]$ and $[SEP]$ marks to the beginning and end of the text respectively.
- **Gloss embeddings**: Gloss texts in WordNet are used to extract word sense embeddings, that is, the vector form of labels. The processing method of these texts is also in accordance with the unified regulations of BERT. Since the entire text needs to be represented, the output vector of $[CLS]$ is used to represent the entire text according to the convention.
- **Example sentence embeddings**: The example sentences in WordNet are used to extract the word sense embeddings, which are another form of word sense definitions. Gloss texts are a conceptual formal description of sense, while example sentences are a description of its applicable scenario. They complement each other, but each has its own focus. In order to extract the scenario information of the target word in the example sentence, we use the $[MASK]$ mark to cover the target word. The vector corresponding to $[MASK]$ in the example sentence is used as the embedding of the example sentence.

At this point, the similarity between the target word W_{target} and each sense S^i can be calculated, where $i \in \{1, ..., |S|\}$ is the index of the list of candidate senses. The list of senses is represented by glosses S^i_{gloss} and example sentences S^i_{exam}. The score of each sense of the target word can be calculated:

$$Score^i_{gloss} = V_{W_{target}} \odot V_{S^i_{gloss}} \tag{16}$$

$$Score^i_{exam} = V_{W_{target}} \odot V_{S^i_{exam}} \tag{17}$$

where V represents the corresponding vectorization (i.e., embedding), and $\odot$ represents the inner product operation.

Quantum Processing Method Under the quantum processing method part, gloss embeddings $V_{S^i_{gloss}}$ and example sentence embeddings $V_{S^i_{exam}}$ are constructed as superposition states,

$$|\mathbf{V}_{S^i_{gloss}}\rangle = \mathcal{F}(V_{S^i_{gloss}}), \tag{18}$$

$$|\mathbf{V}_{S^i_{exam}}\rangle = \mathcal{F}(V_{S^i_{exam}}), \tag{19}$$

and the target word embedding $V_{W_{target}}$ as a measurement operator,

$$M_{W_{target}} = |W_{target}\rangle\langle W_{target}| \tag{20}$$

$$= \mathcal{C}(V_{W_{target}})(\mathcal{C}(V_{W_{target}}))^\dagger. \tag{21}$$

At this point, the score of each sense of the target word can be calculated, i.e., Eq. (15),

$$Q\text{-}Score^i_{gloss} = p(|\mathbf{V}_{S^i_{gloss}}\rangle; M_{W_{target}}), \tag{22}$$

$$Q\text{-}Score^i_{exam} = p(|\mathbf{V}_{S^i_{exam}}\rangle; M_{W_{target}}). \tag{23}$$

Combined with the scores obtained by the classic method, they are collectively used as the parameters of the loss function of QR-WSD.

4.4 Model Training

We employ a cross-entropy loss on the scores of the candidate senses of the target word to train QR-WSD,

$$Score^i = Score^i_{gloss} + Score^i_{exam} + \tag{24}$$

$$Q\text{-}Score^i_{gloss} + Q\text{-}Score^i_{exam}. \tag{25}$$

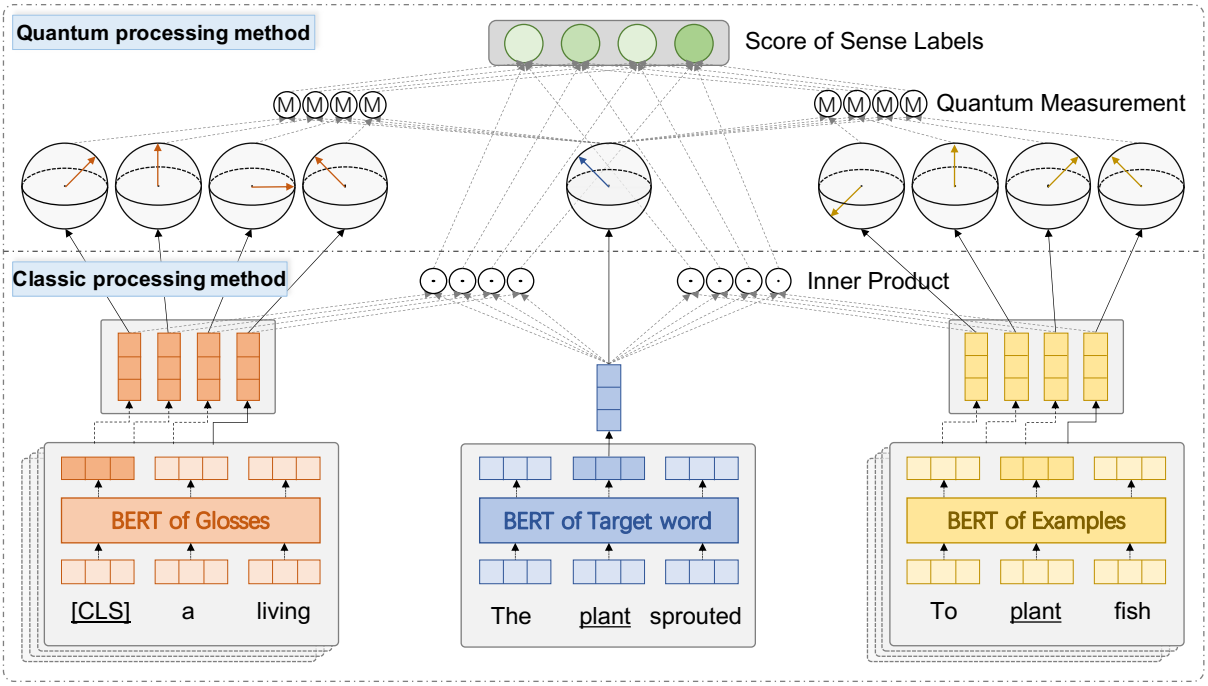

Figure 2: Illustration of model structure. The model consists of two parts: quantum and classical parts. The figure shows three BERTs, but in practice, the BERT of the training target word and the BERT of the training examples are combined into one. $\odot$ represents the inner product operation. The circled M stands for quantum measurement.

The loss function is

$$Loss(Score, index) \tag{26}$$

$$= -\log\left(\frac{\exp(Score^{[index]})}{\sum_{i=1}\exp(Score^{[i]})}\right) \tag{27}$$

$$= -Score^{[index]} + \log\sum_{i=1}\exp(Score^{[i]}) \tag{28}$$

where $index$ is the index of the list of the candidate senses.

QR-WSD employs the Adam optimizer (Kingma and Ba 2015) to update the parameters, and the specific settings of the optimizer will be given in the experimental section.

5 Experiments

5.1 Datasets and Model Settings

Datasets: To evaluate the effectiveness of QR-WSD, we carried out experiments under two evaluation settings, namely **the standardized evaluation setting** and **the enhanced evaluation setting**. The standardized setting includes only SemCor[2] in the training set; the enhanced setting includes SemCor and WNGT[3] in the training set. Both settings use the same development set and test sets. SemEval-07 (**SE7**; Pradhan et al. (2007)), following convention (Kumar et al. 2019; Blevins and Zettlemoyer 2020), is regarded as the development set. Senseval-2 (**SE2**; Edmonds and Cotton (2001)), Senseval-3 (**SE3**; Snyder and Palmer (2004)), SemEval-13 (**SE13**; Navigli, Jurgens, and Vannella (2013)), SemEval-15 (**SE15**; Moro and Navigli (2015)) and their concatenation (**ALL**) are designated as the test sets.

In addition, it needs to be explained that the candidate list of the senses of the target word is all the word senses of WordNet 3.0. Evaluation metrics and other unlisted information are set according to the WSD evaluation framework proposed by Navigli et al. (Navigli, Camacho-Collados, and Raganato 2017).

Model Settings: The computing platform of the program is Ubuntu 18.04, which is equipped with two Tesla P40 GPUs. The program is developed based on the Pytorch 1.8 framework and written in Python 3.6. Moreover, WordNet 3.0 is provided by NLTK 3.5, and *bert-base-uncased* and *bert-large-uncased* are provided by Transformers 4.5. The learning rate, epoch and batch size of the model are $\{1e\text{-}5, 5e\text{-}6\}$, 20 and 4 respectively. Other hyperparameters not listed will be given in the published code.

5.2 Baselines

According to the traditional comparison scheme, we choose two versions of BERT (namely BERT-base and BERT-large) as the encoder to build our model (called QR-WSD$_{\text{base}}$ and QR-WSD$_{\text{large}}$) to compete with the baseline models. The comparison models are divided into models under the standardized setting and under the enhanced setting.

Under the Standardized Setting: Our model is compared with EWISE (Kumar et al. 2019), LMMS (Loureiro and

[2]http://lcl.uniroma1.it/wsdeval/training-data

[3]https://wordnetcode.princeton.edu/glosstag.shtml

Jorge 2019), SREF (Wang and Wang 2020), ARES (Scarlini, Pasini, and Navigli 2020b), BEM (Blevins and Zettlemoyer 2020), EWISER (Bevilacqua and Navigli 2020), SyntagRank (Scozzafava et al. 2020), COF (Wang, Zhang, and Wang 2021b), ESR (Song et al. 2021), Z-Reweighting (Su et al. 2022), and QWSD (Tamburini 2019). Their experimental results come from the values published in the original paper. Among them, QWSD is also a quantum-inspired model, which is our important comparison model.

Under the Enhanced Setting: We choose the three most representative models as the baselines, they are SparseLMMS (Berend 2020), WISER (Bevilacqua and Navigli 2020) and ESR (Song et al. 2021). Their experimental results also come from the values published in the original paper.

5.3 Results and Analysis

Under the Standardized Setting: The experimental results under the standardized evaluation setting are shown in the first block of Tab. 1. From the overall performance point of view, our model is superior to the baseline models in multiple test sets. Compared with QWSD (Tamburini 2019), which is a quantum-inspired model, our model is far superior to the performance of QWSD. A notable reason is that QWSD does not use learnable embeddings, and we choose the current mainstream pre-training language model. Compared with EWISE (Kumar et al. 2019), which is a model in a continuous sense space, our model also performs outstandingly. The reason is that our model not only uses continuous space to constrain embeddings, but also can find better representations through the quantum-inspired representation method.

In terms of overall performance, our results are already in a high position, but still cannot compete with the excellent performance of BEM (Blevins and Zettlemoyer 2020) on multiple test sets, such as *Adj.* and *Adv.*. By analyzing the proportion of head and tail senses in the datasets, it is found that head senses in these datasets have a higher proportion, which does not give full play to the advantages of the quantum-inspired representation.

Under the Enhanced Setting: The experimental results under the enhanced evaluation setting are shown in the second block of Tab. 1. From the experimental results, in addition to the poor performance on the test set *Adj.* and *Adv.*, our model is better than the comparison models under other test sets. By comparing QR-WSD$_{\text{base}}$ and QR-WSD$_{\text{large}}$, we find that based on a more powerful pre-training model will significantly improve the performance.

5.4 Ablation Study under MFSs & LTSs

For a more detailed analysis of the contribution of the quantum-inspired representation to the model, we perform ablation experiments.

In terms of models, we chose the model with the quantum processing method removed as the experimental group, called QR-WSD$^-$, and the original model as the control group, called QR-WSD$^+$. In terms of datasets, the first set of experiments employs the original training set (i.e., SemCor) and test sets; the second set of experiments constructs the Most Frequent Sense (**MFS**) training set and test sets by selecting the head senses in the original training set and test sets respectively; the third set of experiments constructs the Long Tail Sense (**LTS**) training set and test sets by selecting the tail senses in the original training set and test sets respectively. In the implementation process, we classify the senses that appear more than 3 times in the dataset as MFSs, and vice versa as LTSs. Other settings are the same as in the main experiment.

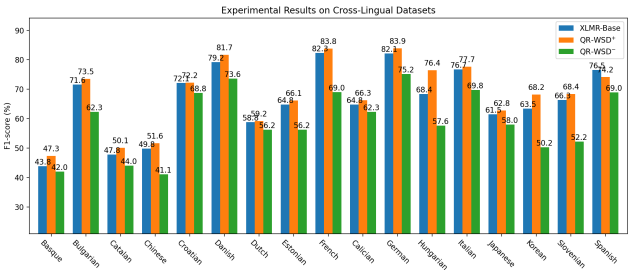

Figure 3: Experimental results of XLMR-Base (which results from data published by the evaluation framework), QR-WSD$^+$, and QR-WSD$^-$ under the cross-lingual WSD evaluation framework.

The experimental results are shown in Tab. 2. Because the datasets used in the first set of experiments are the original data, the results show the performance of the models under the real data. From these results, it can be seen that the quantum representation method has advantages under real data, indicating that the quantum representation has the ability to deal with the head senses and the tail senses at the same time.

Under the MFS datasets, the model with quantum representation does not have a significant advantage. But this result is in line with reality, because when there are only head senses, the classical method can effectively deal with them, and the quantum method is difficult to play its advantages.

Under the LTS datasets, QR-WSD$^+$ is better than QR-WSD$^-$. The reason for presenting this result is that the classical method is not suitable for small sample tasks, while the advantages of the quantum method can be exerted. However, from the overall performance of the experimental results, the quantum representation method is also affected by the number of samples.

5.5 Experiments under Cross-Lingual Datasets

To evaluate the robustness of the model and the applicability of the quantum-inspired representation method, we conduct experiments on the latest cross-lingual evaluation framework[4] proposed by Pasini et al. (Pasini, Raganato, and Navigli 2021).

To clearly demonstrate the effectiveness of the quantum-inspired representation method, we still use the settings of the models in the ablation experiments, namely QR-WSD$^+$

[4]https://sapienzanlp.github.io/xl-wsd/

Models	Dev set	Test sets				Concatenation of all test sets				
	SE7	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	**ALL**
Training data: SemCor										
EWISE (ACL 2019)	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
LMMS (ACL 2019)	68.1	76.3	75.6	75.1	77.0	-	-	-	-	75.4
SREF (EMNLP 2020)	72.1	78.6	76.6	78.0	80.5	80.6	66.5	82.6	84.4	77.8
ARES (EMNLP 2020b)	71.0	78.0	77.1	77.3	83.2	80.6	68.3	80.5	83.5	77.9
BEM (ACL 2020)	74.5	79.4	77.4	79.7	81.7	81.4	68.5	**83.0**	**87.9**	79.0
EWISER (ACL 2020)	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
SyntagRank (ACL 2020)	59.3	71.6	72.0	72.2	75.8	-	-	-	-	71.2
COF (EMNLP 2021b)	69.2	76.0	74.2	78.2	80.9	80.6	61.4	80.5	81.8	76.3
ESR (EMNLP 2021)	**75.4**	80.6	78.2	79.8	82.8	82.5	69.5	82.5	87.3	79.8
Z-Reweighting (ACL; 2022)	71.9	79.6	76.5	78.9	82.5	–	–	–	–	78.6
Quantum-inspired models										
QWSD (RANLP 2019)	-	70.5	69.8	69.8	73.4	73.6	54.4	77.0	80.6	70.6
QR-WSD$_{\text{base}}$	74.5	**80.6**	**79.1**	**80.0**	**84.7**	**83.7**	**71.4**	82.8	86.7	**80.5**
QR-WSD$_{\text{large}}$	75.4	81.8	80.2	81.1	85.1	84.9	72.1	84.4	88.0	81.7
Training data: SemCor + WNGT										
SparseLMMS (EMNLP 2020)	73.0	79.6	77.3	79.4	81.3	-	-	-	-	78.8
EWISER (ACL 2020)	75.2	80.8	**79.0**	80.7	81.8	81.7	66.3	81.2	85.8	80.1
ESR (EMNLP 2021)	77.4	81.4	78.0	81.5	83.9	**83.1**	71.1	**83.6**	**87.5**	80.7
QR-WSD$_{\text{base}}$	**78.1**	**81.5**	**79.0**	**82.0**	**84.4**	**83.1**	**71.2**	80.9	84.1	**81.0**
QR-WSD$_{\text{large}}$	78.8	82.4	81.0	83.4	85.6	84.0	73.0	84.7	87.0	82.1

Table 1: F1-score (%) on the English all-words WSD task. The comparison models are divided into two groups: those under the standardized evaluation setting (i.e., using only SemCor as the training set) and those under the enhanced evaluation setting (i.e., using SemCor and WNGT as the training set). SOTA performance is **bold** compared to QR-WSD$_{\text{base}}$ and underlined compared to QR-WSD$_{\text{large}}$.

	Test sets				
	SE2	SE3	SE13	SE15	**ALL**
Dataset: SemCor					
QR-WSD^{+}	**80.6**	**79.1**	**80.0**	**84.7**	**80.5**
QR-WSD^{-}	77.7	75.3	76.1	82.0	77.1
Dataset: MFS					
QR-WSD^{+}	92.7	**89.3**	**92.0**	**95.8**	**92.4**
QR-WSD^{-}	**93.1**	89.0	91.7	**95.8**	92.2
Dataset: LTS					
QR-WSD^{+}	**63.3**	**67.6**	**71.6**	**71.5**	**70.7**
QR-WSD^{-}	57.1	64.4	63.2	67.7	66.8

Table 2: The experimental results of ablation experiments under the original, MFS and LTS datasets.

and QR-WSD^{-}. The encoders of the model use *bert-base-multilingual-cased*. Furthermore, we adopt the model employed in the evaluation framework as the baseline model, XLMR-Base (Conneau et al. 2020), and the experimental results are also derived from the results published in the paper. Note that since the cross-lingual datasets are constructed based on BabelNet[5], the glosses and example sentences in this section are from BabelNet. Moreover, since most small languages in BabelNet use definitions in English, we directly use glosses and example sentences in English to provide a candidate list of senses.

The experimental results are shown in Fig. 3. Comparing QR-WSD^{+} with XLMR-Base, QR-WSD^{+} is better than XLMR-Base on multiple datasets, which shows that our model has a certain robustness, and also shows that the quantum-inspired representation method has a wide range of applicability. Comparing QR-WSD^{+} with QR-WSD^{-}, QR-WSD^{+} is far superior to QR-WSD^{-}, which shows that quantum representation is helpful for the representation and identification of long-tail senses.

[5]https://babelnet.org/

6 Conclusions

Inspired by the quantum superposition state, this article proposes a novel quantum-inspired representation method. This method attempts to obtain a relatively correct representation for long-tail senses with only a small sample size. Moreover, a WSD model is constructed based on the representation to verify the effect in specific applications. The experimental results show that the model is better than the baseline models and achieves comparable performance. In the long-tail sense dataset constructed and the recently proposed cross-lingual datasets, the quantum-inspired representation shows advantages over the traditional methods, indicating that the representation has a certain potential for solving the long tail phenomenon of data distribution.

The significance of this article is to propose an effective representation method inspired by quantum mechanics, which may explore new ways to solve the data imbalance. In future work, we will further study the application of multiple quantum systems in few-shot tasks, and explore the potential of formalized systems of quantum mechanics in traditional tasks.

References

Abd Elrahman, S. M.; and Abraham, A. 2013. A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing*, 1(2013): 332–340.

Aliwy, A. H.; and Taher, H. A. 2019. Word Sense Disambiguation: Survey Study. *Journal of Computer Science*, 15(7): 1004–1011.

Berend, G. 2020. Sparsity Makes Sense: Word Sense Disambiguation Using Sparse Contextualized Word Representations. In *EMNLP*, 8498–8508.

Bevilacqua, M.; and Navigli, R. 2020. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *ACL*, 2854–2864.

Bevilacqua, M.; Pasini, T.; Raganato, A.; and Navigli, R. 2021. Recent Trends in Word Sense Disambiguation: A Survey. In *IJCAI*, 4330–4338.

Blevins, T.; and Zettlemoyer, L. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *ACL*, 1006–1017.

Bolshina, A.; and Loukachevitch, N. V. 2020. Generating Training Data for Word Sense Disambiguation in Russian. *Proceedings of Conference on Computational linguistics and Intellectual technologies*, 2020: 119–132.

Bruza, P.; Wang, Z.; and Busemeyer, J. R. 2015. Quantum Cognition: A New Theoretical Approach to Psychology. In *Trends in Cognitive Sciences*, volume 19, 383–393.

Busemeyer, J. R.; and Bruza, P. D. 2012. *Quantum Models of Cognition and Decision*. Cambridge University Press.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*, 8440–8451.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Edmonds, P.; and Cotton, S. 2001. SENSEVAL-2: Overview. In *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 1–5.

Fujii, A.; Tokunaga, T.; Inui, K.; and Tanaka, H. 1998. Selective Sampling for Example-Based Word Sense Disambiguation. *Computational Linguistics*, 24(4): 573–597.

Hardy, L. 2001. Quantum Theory From Five Reasonable Axioms. *arXiv: Quantum Physics*.

Holla, N.; Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2020. Learning to Learn to Disambiguate: Meta-Learning for Few-Shot Word Sense Disambiguation. In *EMNLP*, 4517–4533.

Janotta, P.; and Hinrichsen, H. 2014. Generalized probability theories: what determines the structure of quantum theory? *Journal of Physics A*, 47(32): 323001.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kohli, H. 2021. Transfer Learning and Augmentation for Word Sense Disambiguation. In *ECIR*, volume 12657, 303–311.

Kumar, S.; Jat, S.; Saxena, K.; and Talukdar, P. P. 2019. Zero-Shot Word Sense Disambiguation using Sense Definition Embeddings. In *ACL*, 5670–5681.

Li, Q.; Uprety, S.; Wang, B.; and Song, D. 2018. Quantum-Inspired Complex Word Embedding. In *Proceedings of The Third Workshop on Representation Learning for NLP*, 50–57.

Liu, G.; Hou, Y.; and Song, S. 2021. A Quantum-inspired Complex-valued Representation for Encoding Sentiment Information (Student Abstract). In *AAAI*, 15831–15832.

Loureiro, D.; and Jorge, A. M. 2019. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In *ACL*, 5682–5691.

Miller, G. A. 1998. *WordNet: An electronic lexical database*. MIT press.

Moro, A.; and Navigli, R. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, 288–297.

Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM Comput. Surv.*, 41(2): 10:1–10:69.

Navigli, R.; Camacho-Collados, J.; and Raganato, A. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *EACL*, 99–110.

Navigli, R.; Jurgens, D.; and Vannella, D. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics*, 222–231.

Nielsen, M. A.; and Chuang, I. 2002. *Quantum Computation and Quantum Information*. American Association of Physics Teachers.

Ntoutsi, E.; Fafalios, P.; and Gadiraju, U. 2020. Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

Pasini, T.; Raganato, A.; and Navigli, R. 2021. XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In *AAAI*, 13648–13656.

Pradhan, S.; Loper, E.; Dligach, D.; and Palmer, M. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, 87–92.

Scarlini, B.; Pasini, T.; and Navigli, R. 2020a. Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains. In *LREC*, 5905–5911.

Scarlini, B.; Pasini, T.; and Navigli, R. 2020b. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *EMNLP*, 3528–3539.

Scozzafava, F.; Maru, M.; Brignone, F.; Torrisi, G.; and Navigli, R. 2020. Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation. In *ACL*, 37–46.

Snyder, B.; and Palmer, M. 2004. The English all-words task. In *The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 41–43.

Song, Y.; Ong, X. C.; Ng, H. T.; and Lin, Q. 2021. Improved Word Sense Disambiguation with Enhanced Sense Representations. In *EMNLP*, 4311–4320.

Su, Y.; Zhang, H.; Song, Y.; and Zhang, T. 2022. Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting. In *ACL*, 4713–4723.

Tamburini, F. 2019. A Quantum-Like Approach to Word Sense Disambiguation. In *RANLP*, 1176–1185.

Wang, M.; and Wang, Y. 2020. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. In *EMNLP*, 6229–6240.

Wang, M.; and Wang, Y. 2021. Word Sense Disambiguation: Towards Interactive Context Exploitation from Both Word and Sense Perspectives. In *ACL/IJCNLP*, 5218–5229.

Wang, M.; Zhang, J.; and Wang, Y. 2021a. Enhancing the Context Representation in Similarity-based Word Sense Disambiguation. In *EMNLP*, 8965–8973.

Wang, M.; Zhang, J.; and Wang, Y. 2021b. Enhancing the Context Representation in Similarity-based Word Sense Disambiguation. In *EMNLP*, 8965–8973.

Yang, Y.; and Xu, Z. 2020. Rethinking the Value of Labels for Improving Class-Imbalanced Learning. *Advances in neural information processing systems*, 33: 19290–19301.