Quantum Interference Model for Semantic Biases of Glosses in Word Sense Disambiguation

Junwei Zhang[1], **Ruifang He**[2*], **Fengyu Guo**[3], **Chang Liu**[4]
[1]Center for Artificial Intelligence and Intelligent Medicine,
Hangzhou Institute of Medicine, Chinese Academy of Sciences, Zhejiang Province, China.
[2]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China.
[3]College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China.
[4]CSSC Systems Engineering Research Institute, Beijing, China.
zhangjunwei@him.cas.cn, rfhe@tju.edu.cn, fyguo@tjnu.edu.cn, liuc1100101110@163.com

Abstract

Word Sense Disambiguation (WSD) aims to determine the meaning of the target word according to the given context. Currently, a single representation enhanced by glosses from different dictionaries or languages is used to characterize each word sense. By analyzing the similarity between glosses of the same word sense, we find semantic biases among them, revealing that the glosses have their own descriptive perspectives. Therefore, the traditional approach of integrating all glosses by a single representation results in failing to present the unique semantics revealed by the individual glosses. In this paper, a quantum superposition state is employed to formalize the representations of multiple glosses of the same word sense to reveal their distributions. Furthermore, the quantum interference model is leveraged to calculate the probability that the target word belongs to this superposition state. The advantage is that the interference term can be regarded as a confidence level to guide word sense recognition. Finally, experiments are performed under standard WSD evaluation framework and the latest cross-lingual datasets, and the results verify the effectiveness of our model.

Introduction

Word Sense Disambiguation (WSD) aims to determine the word sense of target words according to a given specific context (Bevilacqua et al. 2021; Navigli 2009; Edmonds and Agirre 2006). WSD is a fundamental research topic in the field of Natural Language Processing (NLP), and it has a profound impact on subsequent high-level tasks, such as information retrieval (Zhong and Ng 2012) and machine translation (Parameswarappa 2011).

Accurate and easy-to-distinguish word sense representations can improve the performance of WSD systems (Bevilacqua et al. 2021; Blevins and Zettlemoyer 2020). However, limited by the scarcity of word sense annotations (namely glosses) in dictionaries, obtaining high-quality word sense representations is an important challenge for the WSD task (Scarlini, Pasini, and Navigli 2020b; Kumar et al. 2019; Zhang, He, and Guo 2023; Zhang et al. 2022; Zhang, He, and Guo 2022).

*Corresponding author.

SensEmBERT (Scarlini, Pasini, and Navigli 2020a) and ARES (Scarlini, Pasini, and Navigli 2020b) leverage extensive external knowledge to obtain context-enhanced word sense representations. BEM (Blevins and Zettlemoyer 2020) adopts the bi-encoder joint training method to leverage the context information of the target words to improve the performance of the representations. EWISE (Kumar et al. 2019) replaces the discrete word sense representation space with a continuous space for better generalization performance on both seen and unseen word senses, which mines the inherent constraints between word senses. In addition, methods proposed by Hadiwinoto et al. (Hadiwinoto, Ng, and Gan 2019), Luo et al. (Luo et al. 2018b), Simov et al. (Simov, Osenova, and Popov 2016) are also included.

The above methods employ **a single representation** enhanced by introducing external knowledge or mining intrinsic constraints to describe each word sense, such as glosses from different dictionaries or multilingual glosses. By analyzing the similarity between glosses of the same word sense, we find that there are significant **semantic biases** (Misono et al. 1997; Nikolov and d'Aquin 2020) between the glosses, as shown in Fig. 1. It reveals that it is difficult to effectively coordinate glosses from different definition perspectives (or descriptions from different application scenarios) using a single representation. For example, to characterize a set of features, picking only one of them or taking their average results in inaccurate descriptions and loss of information.

In this paper, we try to use multiple representations to characterize a word sense to simultaneously reveal the distribution of the representations, and then refer to the distribution information to improve the accuracy of word sense recognition. To this end, we adopt the **superposition state** (Nielsen and Chuang 2002) in quantum mechanics to formalize representations from different glosses of the same word sense, and then reveal their distribution. Furthermore, we employ the **quantum interference model** (Busemeyer and Bruza 2012) to calculate the probability that the target word belongs to this superposition state. The advantage of this construction is that the similarity between the target word and each gloss can be calculated, and the interference term derived by the quantum interference model can be

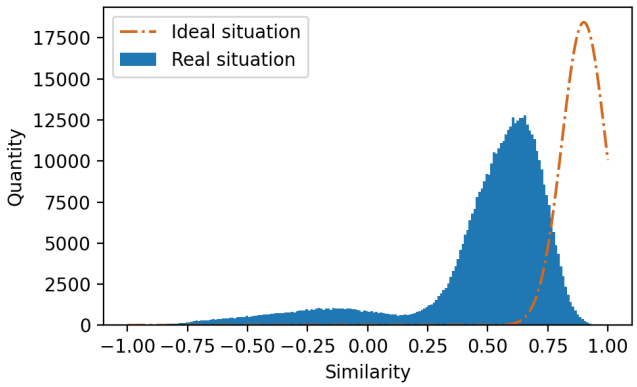

Figure 1: Statistical distribution of similarity between glosses of the same word sense in different dictionaries. Ideally, this distribution should be close to the distribution shown by the dotted line in the figure. See the appendix section for the construction method.

regarded as a confidence level to guide word sense recognition. Specifically, a WSD model is constructed with reference to the physical experiment settings that derive the quantum interference model, in order to train our model to simulate the process of interference experiments. Finally, experiments are performed under standard WSD evaluation framework and the latest cross-lingual datasets, and the results confirm the effectiveness of our model.

The contributions can be summarized as follows:

- Propose to adopt multi-view word sense representation to improve the one-sidedness of single-view representation, and deduce a generalized quantum interference model for the identification of multi-view representation;
- Construct a WSD model with reference to the interference experiment, and achieve comparable performance on a standard WSD evaluation framework and cross-lingual datasets.

Related Work

Word Sense Disambiguation

WSD is a traditional lexical-level task, and its research approaches can be roughly divided into knowledge-based approaches and supervised approaches (Bevilacqua et al. 2021; Navigli 2009; Edmonds and Agirre 2006).

Knowledge-based approaches utilize computational dictionaries, such as WordNet (Miller 1998) or BabelNet (Navigli and Ponzetto 2012), especially their graph structures, to identify contexts in which word senses apply, or to obtain representations of word senses. SyntagRank (Scozzafava et al. 2020) is a purely graph-based method, which uses the Personalized PageRank algorithm to mine the intrinsic relationship between synsets.

Supervised approaches aim to learn a parameterized function that maps from target words to word senses (Bevilacqua et al. 2021). GlossBERT (Huang et al. 2019) is the earliest supervised approach to integrate gloss knowledge. BEM (Blevins and Zettlemoyer 2020) uses gloss knowledge to obtain word sense representations, and leverages the training text to improve the performance of word sense representations through the bi-encoder joint training method. In addition, ARES (Scarlini, Pasini, and Navigli 2020b) obtains context-enhanced word sense representations by augmenting external knowledge; EWISE (Kumar et al. 2019) improves the performance of word sense representations by constraining the word sense space.

The above methods use a single word sense representation to fuse descriptions from different scenarios or glosses from different perspectives, resulting in the obtained representations ignoring the individualized features of each word sense and losing information. Our method uses a multi-view word sense representation to characterize multiple glosses of the same word sense, which not only combats the above shortcomings, but also leverages the semantic biases between glosses to calibrate the decision of word sense recognition.

Quantum-theoretical Approaches

The unique advantages of quantum theory in modeling human cognitive behavior stimulate research interests in quantum cognition (Busemeyer and Bruza 2012; Zhang et al. 2020, 2021a,b). Currently, quantum-theoretical approaches have not received enough attention for the WSD task.

For WSD, models or methods inspired by quantum theory are still in the early stage of research, and they simply use the basic mathematical structures of quantum mechanics. Tamburini (Tamburini 2019), based on quantum probability theory, builds word and sentence embeddings in the complex form and achieves state-of-the-art performance without long training phases. Júnior et al. (Júnior, de Andrade Lopes, and Amancio 2018) and Amancio et al. (Amancio, Oliveira, and da Fontoura Costa 2012) both use complex networks to build WSD systems and the experimental results show that the representation method under the complex form is beneficial to word sense recognition.

Our method uses a quantum superposition state to characterize word sense representations from multiple perspectives, and uses a quantum interference model to obtain output probabilities, so that semantic biases between representations can be used to guide word sense recognition.

Quantum Interference Model

Preliminaries

To understand the quantum interference model and our WSD system, this section introduces the necessary background knowledge of Quantum Probability Theory (QPT). QPT is a generalization of Classical Probability Theory (CPT), and it covers CPT perfectly (Nielsen and Chuang 2002).

Quantum Events: Similar to Kolmogorovian probability theory, QPT assigns probabilities to events; however, the difference is that CPT defines events as sets, while QPT as subspaces of a multidimensional complex Hilbert space $\mathcal{H} = \mathbb{C}^n$.

Quantum States: A quantum state is used to describe a quantum system, and is defined as a complex vector in

Hilbert space using Dirac[1] notation, e.g., $|\psi\rangle \in \mathcal{H}$ with $\||\psi\rangle\| = 1$. A more general high-dimensional quantum state can be defined as

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

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

$$|\Psi\rangle = \phi_1|\psi_1\rangle + \phi_2|\psi_2\rangle + ... + \phi_i|\psi_i\rangle + ... \quad (2)$$

Quantum Measurements: After having a quantum system described by quantum states, the internal situation of the system, that is, the probability distribution of subspaces, can be known through quantum measurement operations (Nielsen and Chuang 2002). Quantum measurements can be defined in various ways, such as general measurement, projection measurement, and POVM measurement (Deutsch 1983). Among them, general measurement is often used in the field of machine learning, so this measurement is mainly introduced.

General measurement is described by a set of measurement operators $\{P_m\}$, where m represents the possible measurement results in the experiment, and these operators act on the state space of the subsystems. When the quantum system is $|\psi\rangle$ before the measurement, then the probability of the measurement result m is defined as

$$\mathcal{P}(m) = \langle\psi|P_m^\dagger P_m|\psi\rangle = \|P_m|\psi\rangle\|^2. \quad (3)$$

After the measurement, the state of the system is changed to

$$|\psi\rangle = |\psi'\rangle = \frac{P_m|\psi\rangle}{\sqrt{\langle\psi|P_m^\dagger P_m|\psi\rangle}}. \quad (4)$$

The measurement operators need to satisfy completeness,

$$\sum_m P_m^\dagger P_m = I, \quad (5)$$

which reveals the fact that the sum of probabilities is 1,

$$\sum_m \mathcal{P}(m) = \sum_m \langle\psi|P_m^\dagger P_m|\psi\rangle = 1. \quad (6)$$

Quantum Interference Experiment

Interference effects are one of the most interesting phenomena that only arise in the field of quantum physics (Jaklevic et al. 1964). The classical interference experiment (also called the double-slit experiment) is often used as a simple example to introduce quantum interference effects, as shown in Fig. 2.

[1]In Dirac notation, $|\cdot\rangle$ is a column vector (called *ket*), while $\langle\cdot|$ is a row vector (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.

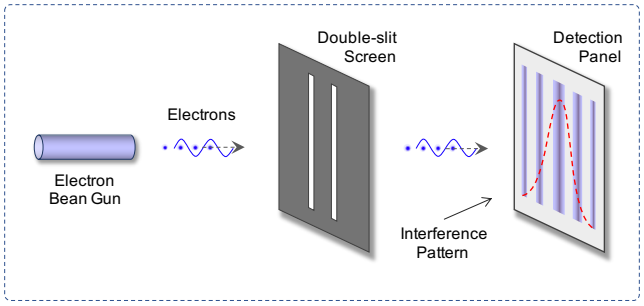

Figure 2: Schematic of the interference experiment.

An electron beam gun fires electrons towards the detection panel through a screen with two slits A and B. The specific position where the electrons land on the detection panel is marked as x. When closing one of the two slits, say B, the probability that the electrons pass through the slit A and land at the position x can be calculated,

$$\mathcal{P}_A(x) = \|P_x|\psi_A\rangle\|^2 \quad (7)$$

where P_x refers to the measurement operator for the position x, and $|\psi_z\rangle$, $z \in \{A, B\}$, refers to the electrons passing through the slit z. Correspondingly, when the slit A is closed, the probability that the electrons pass through the slit B and land at the position x is

$$\mathcal{P}_B(x) = \|P_x|\psi_B\rangle\|^2. \quad (8)$$

Using CPT, the probability that both slits are open and the electrons pass through either slit A or B and land at the position x is

$$\begin{aligned}\mathcal{P}_{AB}(x) &= \mathcal{P}_A(x) + \mathcal{P}_B(x) \\ &= \|P_x|\psi_A\rangle\|^2 + \|P_x|\psi_B\rangle\|^2.\end{aligned} \quad (9)$$

However, the actual experimental results reveal that the double-slit experiment cannot be described by CPT, that is, the above formula Eq. (9) does not hold, and an additional interference term $Int_{AB}(x)$ needs to be added, that is,

$$\mathcal{P}_{AB}(x) = \mathcal{P}_A(x) + \mathcal{P}_B(x) + Int_{AB}(x). \quad (10)$$

In the following, based on QPT, a quantum interference model consistent with the double-slit experimental phenomenon is deduced, and its generalized form is given.

Quantum Interference Model

The double-slit experiment shows that microscopic particles (including electrons) can be in the form of superposition, that is, in the double-slit experiment, an electron passes through the slits A and B at the same time and interferes with each other.

In quantum mechanics, a system in the form of a superposition is described by a superposition state, so the measured system can be defined as

$$|\Psi_{AB}\rangle = \alpha|\psi_A\rangle + \beta|\psi_B\rangle \quad (11)$$

where $\alpha, \beta \in \mathbb{C}$, and $|\alpha|^2 + |\beta|^2 = 1$. When the measurement operator of the observation is denoted as P_x, **the quantum interference model** conforming to the double-slit experimental phenomenon can be defined as

$$\begin{aligned}\mathcal{P}_{AB}(x) =& \|P_x|\Psi_{AB}\rangle\|^2 \\ =& \|P_x(\alpha|\psi_A\rangle + \beta|\psi_B\rangle)\|^2 \\ =& \|\alpha P_x|\psi_A\rangle + \beta P_x|\psi_B\rangle\|^2 \\ =& \|\alpha P_x|\psi_A\rangle\|^2 + \|\beta P_x|\psi_B\rangle\|^2 \\ &+ 2\alpha\beta\cos(\theta)|\langle\psi_A|P_x|\psi_B\rangle|\end{aligned} \tag{12}$$

where θ is the phase angle of the inner product $\langle\psi_A|P_x|\psi_B\rangle$. From this, the interference term in Eq. (10) can be determined,

$$Int_{AB}(x) = 2\alpha\beta\cos(\theta)|\langle\psi_A|P_x|\psi_B\rangle|. \tag{13}$$

Similarly, for high-dimensional superposition states such as Eq. (2), **the generalized form of the quantum interference model** can be defined as

$$\mathcal{P}(x) = \sum_i \|\phi_i P_x|\psi_i\rangle\|^2 + \sum_{i,j;i\neq j} Int_{i,j}(x). \tag{14}$$

Analysis: *Compared with the classical probability model, the quantum interference model adds the interference term, which reveals the impact of the distribution of subsystems on the measurement operator. In specific tasks, the interference term can be given practical meaning. For example, in this paper, the interference term can be regarded as a confidence level (the degree of certainty that the target word belongs to this word sense distribution family) to guide word sense recognition. The interference term derived from the quantum interference model shows the ability to evaluate the distribution of features, which can be analogized to the global perception ability in human cognitive behavior.*

Methodology

Word Sense Disambiguation

WSD belongs to a standard classification task, which aims to learn a mapping from a target word w_{target} given a specific context C to a word sense s_i in a candidate list S of word senses,

$$w_{target} \in C \longrightarrow s_i \in S, \tag{15}$$

where the candidate list of word senses comes from the word senses listed in the dictionary.

Quantum Interference Model for WSD

The physical experimental setup that derives the quantum interference model is itself a cyclic system that repeatedly emits electrons and counts their probability distribution, so building our WSD system (as shown in Fig. 3) with reference to this setup is not only easier to understand but also conforms to the process of organizing components of the quantum interference model.

The electron emitted by the electron bean gun is denoted as P, which has one observation described by a set of measurement operators P_+ and P_-. In the WSD task, this observation is regarded as a word sense observation, so $P_+(P_-)$ can be used to verify the correctness (wrongness) of the word sense of the target word. Due to the completeness of the measurement operators, i.e., Eq. (5), only one of the measurement operators is required. Here, P_+ is selected, that is, we only calculate the probability that this word sense is the correct option.

As shown in Fig. 3, we use the pre-trained model BERT (Devlin et al. 2019) to obtain the word embedding of the target word, V_{target}, transform it into a quantum state by borrowing the sum of squares normalization function,

$$\begin{aligned}|\psi_{target}\rangle &= \mathcal{SSN}(V_{target}) \\ &= \frac{1}{\sqrt{\sum_i V_{target,i}^2}} V_{target},\end{aligned} \tag{16}$$

and then construct it as the measurement operator of the word sense observation of the target word,

$$P_+^{target} = |\psi_{target}\rangle\langle\psi_{target}|. \tag{17}$$

After passing through the screen, the electron is modified into a superposition state $|\Psi\rangle$ under a set of quantum states $\{|\psi_k\rangle\}_{k=1}^K$. In the WSD task, the set of quantum states is different for different word senses of the target word. Assuming that there are three different glosses for the word sense $sense_i$, i.e., $K = 3$, the set of quantum states is denoted as $\{|\psi_{gloss_k}^{sense_i}\rangle\}_{k=1}^{K=3}$.

As shown in Fig. 3, we use the pre-trained model BERT to obtain the text embeddings of the glosses of the word sense $sense_i$, $V_{gloss_k}^{sense_i}$, transform them into the quantum states by borrowing the sum of squares normalization function,

$$|\psi_{gloss_k}^{sense_i}\rangle = \mathcal{SSN}(V_{gloss_k}^{sense_i}), \tag{18}$$

and then construct them into a superposition state

$$|\Psi^{sense_i}\rangle = \sum_{k=1}^{K} \phi_k |\psi_{gloss_k}^{sense_i}\rangle \tag{19}$$

where ϕ_k can be set to a specific value, or can be obtained by Neural Network based on V_{target} and $V_{gloss_k}^{sense_i}$,

$$\phi_k = \mathcal{NN}([V_{target}, V_{gloss_s}^k]). \tag{20}$$

After landing on the detection panel, the electron is counted by the corresponding observations, and the probability of the measurement operators of the observations can be calculated using the quantum interference model, i.e., Eq. (14). In the WSD task, the probability value obtained by the measurement operator P_+^{target} and the superposition state $|\Psi^{sense_i}\rangle$ through the quantum interference model,

$$\begin{aligned}\mathcal{P}_{sense_i}(+) = &\sum_{k=1}^{K} \|\phi_k P_+^{target} |\psi_{gloss_k}^{sense_i}\rangle\|^2 \\ &+ \sum_{k,k';k\neq k'} Int_{gloss_k,gloss_{k'}}^{sense_i}(+),\end{aligned} \tag{21}$$

can be regarded as the similarity between the word vector of the target word and the text vector of the word sense $sense_i$. Therefore, this value can be used to finally determine the correct word sense of the target word.

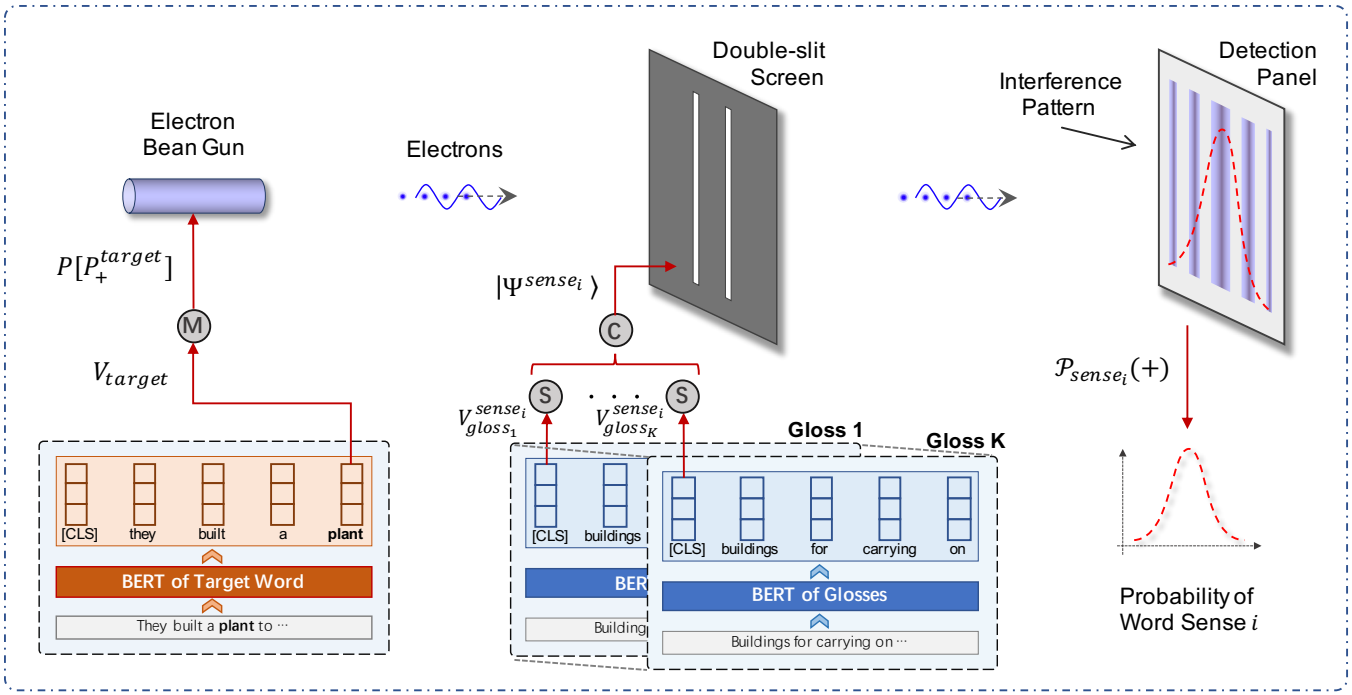

Figure 3: Schematic diagram of the model structure. The circled M, S and C refer to constructing measurement operators, generating quantum states and combining superposition states, respectively.

Model Training

The quantum interference experiment is essentially to repeatedly emit electrons and measure the electrons to obtain their probability distribution. Therefore, the training process of our WSD system can be regarded as cyclically identifying the word sense of the target words in the training set and updating the parameters of the quantum interference model through the loss function.

The loss function of the system is defined as

$$\begin{aligned} & Loss(Score, index) \\ & = -\log\left(\frac{\exp(Score_{[index]})}{\sum_{i=1}\exp(Score_{[i]})}\right) \\ & = -Score_{[index]} + \log\sum_{i=1}\exp(Score_{[i]}) \end{aligned} \tag{22}$$

where $index$ is the index of the candidate list of word senses of the target word and

$$Score = [\mathcal{P}_{sense_1}(+), ..., \mathcal{P}_{sense_i}(+), ...]. \tag{23}$$

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

Experiments

To answer the following three questions through experimental analysis:

- **Q1:** Compared to baseline models and previous work, does the WSD model constructed from the quantum interference model have comparable advantages?
- **Q2:** Do the multi-perspective word sense representation and the interference term derived from the quantum interference model play a role in word sense recognition?
- **Q3:** Can the model be extended to other languages, that is, what is its generalization performance?

Datasets

Our WSD model is evaluated under the standard English all-words WSD evaluation framework proposed by Navigli et al. (Navigli, Camacho-Collados, and Raganato 2017). The training set is SemCor[2] (Miller et al. 1993); the development set is selected as SemEval-2007 (SE07; Pradhan et al. (2007)) by convention; the test sets include Senseval-2 (SE2; Palmer et al. (2001)), Senseval-3 (SE3; Snyder and Palmer (2004)), SemEval-2013 (SE13; Navigli, Jurgens, and Vannella (2013)), SemEval-2015 (SE15; Moro and Navigli (2015)), and the combination of all test sets (shown as *ALL*). In addition, the sets of nouns, verbs, adjectives and adverbs extracted from *ALL* are also used as the test set. By convention, F1-score in percentage is used as an evaluation metric. All glosses used in this paper are from BabelNet[3].

Baselines

To evaluate our model and determine its place in the WSD community, we divide the comparison models into two groups, namely the previous work and the baseline systems.

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

[3]https://babelnet.org/

Models	Dev set	Test sets				Concatenation of all test sets				
	SE07	SE2	SE3	SE13	SE15	*N.*	*V.*	*Adj.*	*Adv.*	*ALL*
Previous Work										
HCAN (EMNLP; 2018a)	–	72.8	70.3	68.5	72.8	72.7	58.2	77.4	84.1	71.1
GAS (ACL; 2018c)	–	72.2	70.5	67.2	72.6	72.2	57.7	76.6	85.0	70.6
GLU (EMNLP; 2019)	68.1	75.5	73.6	71.1	76.2	–	–	–	–	74.1
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
SyntagRank (ACL; 2020)	59.3	71.6	72.0	72.2	75.8	–	–	–	–	71.2
COF (EMNLP; 2021)	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
Baseline Systems										
GlossBERT (EMNLP; 2019)	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
EWISE (ACL; 2019)	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
BEM (ACL; 2020)	74.5	79.4	77.4	79.7	81.7	81.4	68.5	**83.0**	***87.9***	79.0
SensEmBERT (AAAI; 2020a)	73.6	80.6	70.3	74.8	80.2	–	–	–	–	75.9
ARES (EMNLP; 2020b)	71.0	78.0	77.1	77.3	***83.2***	80.6	68.3	80.5	83.5	77.9
Quantum-inspired models										
QWSD (RANLP; 2019)	–	70.5	69.8	69.8	73.4	73.6	54.4	77.0	80.6	70.6
Our$_{\text{-Base}}$	74.4	**80.8**	**78.5**	**80.0**	82.9	**82.7**	**69.5**	**83.0**	86.5	**80.1**
Our$_{\text{-Large}}$	75.0	***81.5***	***79.2***	***81.2***	***83.2***	***84.4***	***71.7***	***84.4***	87.5	***81.6***

Table 1: F1-score (%) on the English all-words WSD task. The experimental results are divided into two parts: one is the previous work, and the other is the baseline systems. The best performance compared to Our$_{\text{-Base}}$ and Our$_{\text{-Large}}$ is shown in bold and bold italics, respectively.

For the previous work, we select excellent models from the past five years. HCAN (Luo et al. 2018a) and GAS (Luo et al. 2018c) in 2018, LMMS (Loureiro and Jorge 2019) and GLU (Hadiwinoto, Ng, and Gan 2019) in 2019, SREF (Wang and Wang 2020), ARES (Scarlini, Pasini, and Navigli 2020b) and SyntagRank (Scozzafava et al. 2020) in 2020, COF (Wang, Zhang, and Wang 2021) and ESR (Song et al. 2021) in 2021, and Z-Reweighting (Su et al. 2022) in 2022 are selected. Their results are taken from the published data of the original paper.

For the baseline systems, we compare GlossBERT (Huang et al. 2019) which first integrates gloss knowledge, EWISE (Kumar et al. 2019) which employs continuous space to constrain word sense representations, BEM (Blevins and Zettlemoyer 2020) which borrows training text to improve word sense representations, and SensEmBERT (Scarlini, Pasini, and Navigli 2020a) and ARES (Scarlini, Pasini, and Navigli 2020b) which strengthen word sense representations by introducing external knowledge. In addition, we compare QWSD (Tamburini 2019) which is also inspired by quantum theory. Their results are also taken from the original paper.

Settings

The operating platform of the hardware is Ubuntu 18.04.3 and has six GPU whose version is NVIDIA GeForce RTX 3090. The development platform is Python 3.8.3[4], and the learning framework is Pytorch 1.8.1[5]. The pretrained language model BERT (Devlin et al. 2019) is provided by Transformers 4.5.1[6]. Following the traditional comparison method, the versions *BERT-base-uncased* and *BERT-large-uncased* are used to build the encoders of our model, and the constructed model are called **Our**$_{\text{-Base}}$ and **Our**$_{\text{-Large}}$, respectively. The hyperparameters *Learning Rate*, *Context Batch Size*, *Gloss Batch Size*, *Epochs*, *Context Maximum Length* and *Gloss Maximum Length* of the model are set to $[1E\text{-}5, 5E\text{-}6, 1E\text{-}6]$, $[1, 2, 3, 4]$, 256, 20, 128 and 32, respectively. Hyperparameters not listed are given in the published code.

Results and Analysis

For **Q1**, the following experimental analysis provides an answer. The experimental results under the English all-words WSD evaluation framework are shown in Tab. 1.

Our model performs well on multiple test sets, indicating that the scheme of characterizing multi-view word sense representation based on the quantum superposition state and using the quantum interference model to calculate the probability is effective. In addition, both **Our**$_{\text{-Base}}$ and **Our**$_{\text{-Large}}$ have excellent performance on the important test set *ALL*, which proves that our model is reliable.

Compared with previous work, the gap opened by **Our**$_{\text{-Base}}$ is not significant because they are not based on the same experimental settings and using the same external re-

[4]https://www.python.org/

[5]https://pytorch.org/

[6]https://huggingface.co/

Models	Test sets				
	SE2	SE3	SE13	SE15	*ALL*
Our$^{\text{-single}}_{\text{-Base}}$	79.4	77.4	79.7	81.7	79.0
Our$^{\text{-inter}}_{\text{-Base}}$	77.8	75.5	76.5	80.0	77.1
Our$_{\text{-Base}}$	**80.8**	**78.5**	**80.0**	**82.9**	**80.1**

Table 2: The experimental results of ablation study under the English all-words WSD evaluation framework.

sources; **Our**$_{\text{-Large}}$ achieves outstanding performance, indicating that relying on more powerful pre-trained models can further stimulate potential.

Compared with the baseline systems, the experimental results prove that the multi-view word sense representation method proposed in this paper is definitely helpful to improve the performance. Comparing similar quantum-inspired model QWSD (Tamburini 2019), it turns out that representations based on quantum superposition states are better than representations based on complex numbers.

Analysis of poor results: On the training set, our model is inferior to ESR (Song et al. 2021), but better than the other comparison models, possibly because our model does not overfit the training set. On the test set SE15, our model **Our**$_{\text{-Base}}$ is inferior to ARES (Scarlini, Pasini, and Navigli 2020b). Through the analysis of this dataset, it is found that the dataset mostly uses high-frequency word senses. Because the recognition difficulty of this dataset is relatively low, the multi-view word-sense representation of our model will actually drag down the overall performance. By analyzing the constructed test set *Adv.*, we find the same phenomenon as SE15. This result once again proves that for a dataset with a large proportion of high-frequency word senses, choosing a reliable single representation is the best choice.

Ablation Study

For **Q2**, ablation experiments are used to answer the question. We replace the original multi-view word sense representation with a single representation to construct the ablation model **Our**$^{\text{-single}}_{\text{-Base}}$, and delete the interference term on the original model to construct the ablation model **Our**$^{\text{-inter}}_{\text{-Base}}$. The experiments are carried out under the standard English all-words WSD evaluation framework, and the experimental results are shown in Tab. 2.

From Tab. 2, our improvement is worthwhile. Compared with **Our**$^{\text{-single}}_{\text{-Base}}$, the original model has an improvement of more than one percentage point on each test set, indicating that the multi-view word sense representation does help to improve the performance of word sense recognition. Compared with **Our**$^{\text{-inter}}_{\text{-Base}}$, the original model has an improvement of three percentage points on multiple test sets, revealing that the quantum interference term can serve as a meaningful supervisory signal to improve the performance of word sense recognition, and indirectly indicating that characterizing the distribution of word senses is valuable.

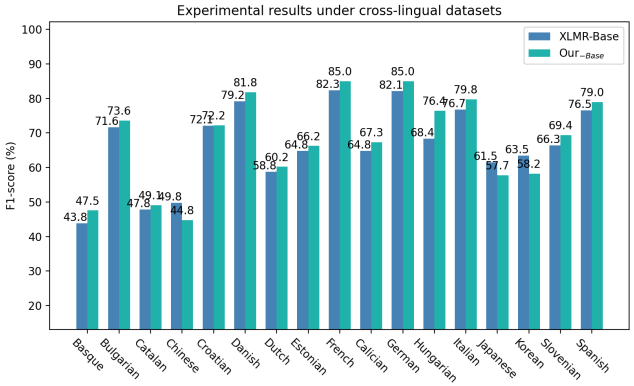

Figure 4: The experimental results under the cross-lingual datasets.

Experiments on Cross-Lingual Datasets

We perform experiments under the cross-lingual datasets to answer **Q3**. The datasets[7] are proposed by Pasini et al. (Pasini, Raganato, and Navigli 2021) Since it is difficult to obtain glosses in the corresponding language, here we adopt a compromise solution to replace glosses in other languages with English glosses. The pretrained language model used in our model is also adjusted to *bert-base-multilingual-cased* accordingly. The comparison model selects XLMR-Base (Conneau et al. 2020) provided in the original paper of the datasets, and the experimental results are shown in Fig. 4.

From Fig. 4, our model outperforms the comparison model in most languages, indicating that our model has certain robustness. By analyzing the underperforming datasets, we find that they are mostly niche or non-alphabetic languages. The reason for the poor results is that we use English glosses instead of glosses in other languages, so languages with similar character types to English will perform better.

Conclusions

Aiming at the problem that the current WSD models leverage external knowledge from different sources to enhance a single word sense representation, this paper proposes a multi-view word sense representation to deal with the semantic biases existing in external knowledge. We adopt a quantum superposition state to formalize glosses from different sources, and employ the quantum interference model to obtain the probability that the target word belongs to this superposition state. The advantage of this construction is that the interference term derived from the quantum interference model can be regarded as a confidence level (that is, a supervisory signal from the gloss distribution information) to improve word sense recognition. The contribution of this paper is to propose an effective alternative to the traditional single word-sense representation, and to verify the validity of this representation on extensive experiments.

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

Acknowledgments

Our work is supported by the National Natural Science Foundation of China (No. 61976154, 62376192, 62376188 and 62106176) and the National Key Research and Development Program of China (No. 2022YFC3600902).

References

Amancio, D. R.; Oliveira, O. N.; and da Fontoura Costa, L. 2012. Unveiling the relationship between complex networks metrics and word senses. *EPL (Europhysics Letters)*, 98(1): 18002.

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*, 5670–5681.

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.

Deutsch, D. 1983. Uncertainty in quantum measurements. *Physical Review Letters*, 50(9): 631.

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 Agirre, E. 2006. *Word Sense Disambiguation: Algorithms And Applications*. Springer Verlag. Text, Speech and Language Technology Series.

Hadiwinoto, C.; Ng, H.; and Gan, W. C. 2019. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. In *EMNLP/IJCNLP*, 5297–5306.

Huang, L.; Sun, C.; Qiu, X.; and Huang, X. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *EMNLP*, 3507–3512.

Jaklevic, R.; Lambe, J.; Silver, A.; and Mercereau, J. 1964. Quantum interference effects in Josephson tunneling. *Physical Review Letters*, 12(7): 159.

Júnior, E. A. C.; de Andrade Lopes, A.; and Amancio, D. R. 2018. Word sense disambiguation: A complex network approach. *Inf. Sci.*, 442: 103–113.

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

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

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.

Luo, F.; Liu, T.; He, Z.; Xia, Q.; Sui, Z.; and Chang, B. 2018a. Leveraging Gloss Knowledge in Neural Word Sense Disambiguation by Hierarchical Co-Attention. In *EMNLP*, 1402–1411.

Luo, F.; Liu, T.; Xia, Q.; Chang, B.; and Sui, Z. 2018b. Incorporating Glosses into Neural Word Sense Disambiguation. In *ACL*, 2473–2482.

Luo, F.; Liu, T.; Xia, Q.; Chang, B.; and Sui, Z. 2018c. Incorporating Glosses into Neural Word Sense Disambiguation. In *ACL*, 2473–2482.

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

Miller, G. A.; Leacock, C.; Tengi, R.; and Bunker, R. 1993. A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*, 303–308.

Misono, Y.; Mazuka, R.; Kondo, T.; and Kiritani, S. 1997. Effects and Limitations of Prosodic and Semantic Biases on Syntactic Disambiguation. *Journal of Psycholinguistic Research*, 26: 229–245.

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.

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

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

Nikolov, A.; and d'Aquin, M. 2020. Uncovering Semantic Bias in Neural Network Models Using a Knowledge Graph. In *CIKM*, 1175–1184.

Palmer, M.; Fellbaum, C. D.; Cotton, S.; Delfs, L.; and Dang, H. T. 2001. English Tasks: All-Words and Verb Lexical Sample. In *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 21–24.

Parameswarappa, S. B. 2011. Kannada Word Sense Disambiguation for Machine Translation. *International Journal of Computer Applications*, 34: 1–8.

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. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *AAAI*, 8758–8765.

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.

Simov, K. I.; Osenova, P. N.; and Popov, A. 2016. Using Context Information for Knowledge-Based Word Sense Disambiguation. In *AIMSA*, volume 9883, 130–139.

Snyder, B.; and Palmer, M. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, 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.; Zhang, J.; and Wang, Y. 2021. Enhancing the Context Representation in Similarity-based Word Sense Disambiguation. In *EMNLP*, 8965–8973.

Zhang, J.; He, R.; and Guo, F. 2022. Bi-matching Mechanism to Combat Long-tail Senses of Word Sense Disambiguation. In *ECML-PKDD*, volume 13714, 621–637.

Zhang, J.; He, R.; and Guo, F. 2023. Quantum-Inspired Representation for Long-Tail Senses of Word Sense Disambiguation. In *AAAI*, 13949–13957.

Zhang, J.; He, R.; Guo, F.; Ma, J.; and Xiao, M. 2022. Disentangled Representation for Long-tail Senses of Word Sense Disambiguation. In *CIKM*, 2569–2579.

Zhang, J.; He, R.; Li, Z.; Zhang, J.; Wang, B.; Li, Z.; and Niu, T. 2021a. Quantum Correlation Revealed by Bell State for Classification Tasks. In *IJCNN*, 1–8.

Zhang, J.; Hou, Y.; Li, Z.; Zhang, L.; and Chen, X. 2020. Strong Statistical Correlation Revealed by Quantum Entanglement for Supervised Learning. In *ECAI*, volume 325, 1650–1657.

Zhang, J.; Li, Z.; He, R.; Zhang, J.; Wang, B.; Li, Z.; and Niu, T. 2021b. Interactive Quantum Classifier Inspired by Quantum Open System Theory. In *IJCNN*, 1–7.

Zhong, Z.; and Ng, H. T. 2012. Word Sense Disambiguation Improves Information Retrieval. In *ACL*, 273–282.

Construction method of Fig. 1

For Fig. 1, the statistics come from multiple glosses for the same word sense in BabelNet 5.1 (mainly using the two glosses from WordNet 3.0 and Wikipedia collected by BabelNet). The original embeddings of the glosses are obtained from the original pre-trained language model BERT, and the ideal embeddings are obtained from the trained BERT model (that is, the optimized and learned BERT model in this paper). The calculation method of similarity between the embeddings of two glosses of the same word sense includes: measuring the Euclidean distance between vectors, and setting the similarity of mutually orthogonal vectors to 0, and constraining other values to be between -1 and 1. Finally, the statistical results of the original glosses and the ideal glosses, that is, the distribution of the real situation and the distribution of the ideal situation, are obtained respectively. The ideal distribution is smoothed.