Quantum-inspired Non-homologous Representation Constraint Mechanism for Long-tail Senses of Word Sense Disambiguation

Junwei Zhang, Xiaolin Li[*]

Hangzhou Institute of Medicine, Chinese Academy of Sciences, Hangzhou, Zhejiang, China
zhangjunwei@him.cas.cn, xiaolinli@ieee.org

Abstract

Word Sense Disambiguation (WSD) aims to determine the meaning of target words according to the given context. The recognition of high-frequency senses has reached expectations, and the current research focus is mainly on low-frequency senses, namely Long-tail Senses (LTSs). One of the challenges in long-tail WSD is to obtain clear and distinguishable definition representations based on limited word sense definitions. Researchers try to mine word sense definition information from data from different sources to enhance the representations. Inspired by quantum theory, this paper provides a constraint mechanism for representations under non-homogeneous data to leverage the geometric relationship in its Hilbert space to constrain the value range of parameters, thereby alleviating the dependence on big data and improving the accuracy of representations. We theoretically analyze the feasibility of the constraint mechanism, and verify the WSD system based on this mechanism on the standard evaluation framework, constructed LTS datasets and cross-lingual datasets. Experimental results demonstrate the effectiveness of the scheme and achieve competitive performance.

Introduction

Word Sense Disambiguation (WSD) aims to determine the meaning of target words according to the given context, which belongs to the basic research topic in the field of natural language processing (Bevilacqua et al. 2021; Navigli 2009). The accuracy of WSD is of great significance and value to downstream tasks (Kaddoura, Ahmed, and D. 2022), such as machine translation (Campolungo et al. 2022), information retrieval (Abderrahim and Abderrahim 2022), sentiment analysis (Farooq et al. 2015), etc. In addition, WSD, like text classification, is often used as a training ground for new models and methods.

WSD systems have reached expectations for the recognition of high-frequency senses (that is, commonly used word senses), and the current research focus is mainly on the recognition of low-frequency senses (that is, rarely used word senses), also known as Long-Tail Senses (LTSs) (Su et al. 2022; Zhang et al. 2022b; Chen, Zhang, and He 2022; Du et al. 2021; Blevins and Zettlemoyer 2020). The root of the difficulty in long-tail WSD lies in:

- LTSs are rarely used, which leads to insufficient training samples;
- LTSs lack clear and distinguishable definitions, which leads to inaccurate representation of word sense definitions.

Among them, training samples refer to the application scenarios of word senses, such as example sentences given in dictionaries, which provide the context information of target words; word sense definitions refer to the descriptions of word senses, such as glosses given in dictionaries, which provide word sense label information. In the process of WSD, the two jointly affect the accuracy of word sense recognition. There are many methods to deal with the lack of training samples of LTSs, and the solution is similar to the classic few-shot task, so this paper focuses on improving the inaccurate representation of word sense definitions.

Kumar et al. (Kumar et al. 2019) constructed word sense definition representations in a continuous space to constrain unclear representations through clear ones. Blevins et al. (Blevins and Zettlemoyer 2020) leveraged the semantic knowledge in the training samples to enhance the word sense definition representations through the joint training of dual encoders (that is, the target word encoder and the definition encoder). Zhang et al. (Zhang et al. 2022a) realized the enhancement of word sense definition representations by integrating example sentences and definitions of related word senses in WordNet. It is not difficult to see from the relevant research in recent years that the effective means of strengthening word sense definition representations is to integrate external knowledge and impose constraint mechanisms.

Inspired by quantum theory (Nielsen and Chuang 2002), this paper proposes a quantum-like model that can simultaneously expand external knowledge and enforce constraints. More importantly, the constraint mechanism not only provide constraints in continuous space, but also implement geometric constraints in Hilbert space for representations of non-homogeneous data. Specifically, firstly, we map homologous or non-homologous data into corresponding embeddings through classical language models, such as BERT (Devlin et al. 2019); secondly, impose normalization constraints on the embeddings to make them quantum states; thirdly, combine the states into a superposition state; finally, the quantum measurement operation is implemented to obtain the result.

[*] Corresponding author.

Note that homologous data refers to data generated by the same generator (that is, has the same structure and obeys similar statistical laws); otherwise, it is non-homologous data. The quantum-like model exhibits spatial geometry constraints only when the generated representations come from non-homogeneous data. If the model is used on homogeneous data, it can only provide representation combinations and continuous space constraints.

Our contributions can be summarized as follows:

- A quantum-like model that can simultaneously provide representation combinations, continuous space constraints, and spatial geometry constraints is proposed, inspired by quantum theory;
- A theoretical analysis of why the geometric constraint mechanism can be helpful for few-shot learning tasks is given;
- Finally, the WSD system based on the quantum-like model is verified on the standard evaluation framework, constructed LTS datasets and cross-lingual datasets; the experimental results prove the effectiveness of the quantum-like model and achieve state-of-the-art performance.

Related Work

Long-tail WSD

The recognition process of WSD is a matching process between the target word embedding and the text embedding of word sense definitions, so the solutions for long-tail WSD can be roughly divided into two categories, enhancing the representation of target words and improving the representation of word sense definitions. The research focus of this paper is on improving the representation of word sense definitions.

Huang et al. (Huang et al. 2019) first proposed to train the representation of word sense labels through the word sense definitions in the dictionary. Subsequently, Blevins et al. (Blevins and Zettlemoyer 2020) enhanced the representation of word sense definitions through the semantic information in the training samples; Yap et al. (Yap, Koh, and Chng 2020) and Zhang et al. (Zhang, He, and Guo 2021) borrowed the example sentences from the dictionary; Scarlini et al. (Scarlini, Pasini, and Navigli 2020a) used multilingual knowledge. Kumar et al. (Kumar et al. 2019) leveraged continuous space constraints to enhance unclear representations from clear representations of word sense definitions. In addition, Kumar et al. (Kumar et al. 2019) and Holla et al. (Holla et al. 2020) also transferred solutions from the fields of few-shot learning and meta learning.

It is not difficult to see from the related work in recent years that integrating external resources and imposing constraint mechanisms are the mainstream methods to deal with long-tail WSD. The work in this paper attempts to achieve the above two functions in one model, and provides spatial geometry constraints for representations under non-homogeneous data.

Quantum-inspired Models for WSD

Quantum-inspired models refer to learning models based on quantum theory (Nielsen and Chuang 2002) or quantum cognition (Busemeyer and Bruza 2012). Since the success of the quantum language model proposed by Basile et al. (Basile and Tamburini 2017), the intersection of quantum theory and natural language processing has gradually become a research hotspot (Li et al. 2016; Xie et al. 2015).

In the field of WSD, related research is still in the early stages of exploration. The model QWSD proposed by Tamburini et al. (Tamburini 2019) is the first of its kind. Although QWSD is very simple in model design, its advantage is that it does not require a long training process. In addition, the WSD system proposed by Zhang et al. (Zhang et al. 2022b) uses the disentanglement representation inspired by quantum theory.

Our work also leverages the mathematical form of quantum theory, namely quantum probability theory, to realize that the constructed WSD model can combat the data sparsity problem faced by long-tail WSD.

Theoretical Analysis

Preliminaries

Before analyzing the constraint mechanism and presenting the quantum-like model, the necessary preliminaries of Quantum Probability Theory (QPT) is given. QPT is a more general probability theory, which is backward compatible with Classical Probability Theory (CPT), namely Kolmogorov probability theory. QPT, like CPT, can be used as a modeling tool for information systems. See Ref. (Nielsen and Chuang 2002) for more details.

Quantum Events: QPT assigns probabilities to events like CPT, but the difference is that events in quantum probability are described by subspaces in Hilbert space $\mathcal{H} \in \mathbb{C}^n$, while events in classical probability are described by sets.

Quantum States: Quantum states, also called quantum systems, are described as complex vectors $|\psi\rangle \in \mathcal{H}$ in Hilbert space using the Dirac[1] notation. Quantum superposition states refer to quantum states in which multiple states are superimposed at the same time. Its formalization is defined as

$$|\psi\rangle = \varepsilon_1|e_1\rangle + \varepsilon_2|e_2\rangle + ... + \varepsilon_i|e_i\rangle + ... \tag{1}$$

where ε_i is called the probability amplitude, $\varepsilon_i = \langle e_i|\psi\rangle \in \mathbb{C}$, $\sum_i |\varepsilon_i|^2 = 1$, and $|e_i\rangle$ is the basis state of $\mathcal{H}$. In general, superposition states can also be composed of other superposition states,

$$|\Psi\rangle = \varepsilon_1|\psi_1\rangle + \varepsilon_2|\psi_2\rangle + ... + \varepsilon_i|\psi_i\rangle + ... \tag{2}$$

Quantum Measurements: The mainstream quantum measurement methods include general measurement, projection measurement and POVM measurement. Among

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

them, general measurement is often used in the field of machine learning, so only general measurement is presented here.

General measurement is described by a set of measurement operators $\{M_m\}$, and they satisfy completeness,

$$\sum_m M_m^\dagger M_m = I, \tag{3}$$

where m refers to a possible measurement result in the experiment, and I refers to the identity matrix. The quantum system is in $|\psi\rangle$ before being measured, and the probability of the possible result m is

$$p(m)=p(M_m;|\psi\rangle)=\langle\psi|M_m^\dagger M_m|\psi\rangle = \|M_m|\psi\rangle\|^2; \tag{4}$$

after being measured, the quantum system is in

$$|\psi'\rangle = \frac{M_m|\psi\rangle}{\sqrt{\langle\psi|M_m^\dagger M_m|\psi\rangle}}. \tag{5}$$

From the completeness of the measurement operators, it can be deduced that

$$\sum_m p(m) = \sum_m \langle\psi|M_m^\dagger M_m|\psi\rangle = 1. \tag{6}$$

Theoretical Analysis for Quantum-inspired Constraint Mechanism

In the field of quantum information processing, the characteristic representations from data are described as quantum states, multi-source data are often integrated as quantum superposition states, and possible results are described by measurement operators (Nielsen and Chuang 2002; Zhang et al. 2020, 2021, 2022c,d).

Accordingly, the quantum states $|\psi_A\rangle$ and $|\psi_B\rangle$ obtained from non-homologous data, say A and B, can be constructed as the superposition state,

$$|\Psi_{AB}\rangle = \alpha|\psi_A\rangle + \beta|\psi_B\rangle, \tag{7}$$

where α, $\beta \in \mathbb{R}$ are the probability amplitudes of constructing the superposition state, and $\alpha^2 + \beta^2 = 1$. In specific tasks, it can be a superposition state composed of multiple quantum states, or a probability amplitude in the form of complex numbers. The measurement operators corresponding to the possible results can be described by a concrete basis state $|e_i\rangle$, $M_m = |e_i\rangle\langle e_i|$, or by a quantum state $|\phi\rangle$ obtained from the label data ϕ,

$$M_m = |\phi\rangle\langle\phi|. \tag{8}$$

The probability of the possible result m is

$$\begin{aligned} p(M_m;|\Psi_{AB}\rangle) &= \|M_m|\Psi_{AB}\rangle\|^2 \qquad (9)\\ &= \|M_m(\alpha|\psi_A\rangle + \beta|\psi_B\rangle)\|^2 \\ &= \|\alpha M_m|\psi_A\rangle + \beta M_m|\psi_B\rangle\|^2 \\ &= \|\alpha M_m|\psi_A\rangle\|^2 + \|\beta M_m|\psi_B\rangle\|^2 + Int, \end{aligned}$$

where the interference term Int is

$$\begin{aligned} Int &= 2\alpha\beta\cos(\theta)|\langle\psi_A|M_m|\psi_B\rangle| \qquad (10)\\ &= 2\alpha\beta\cos(\theta)|\langle\psi_A|\phi\rangle\langle\phi|\psi_B\rangle| \end{aligned}$$

and θ is the phase angle between quantum states.

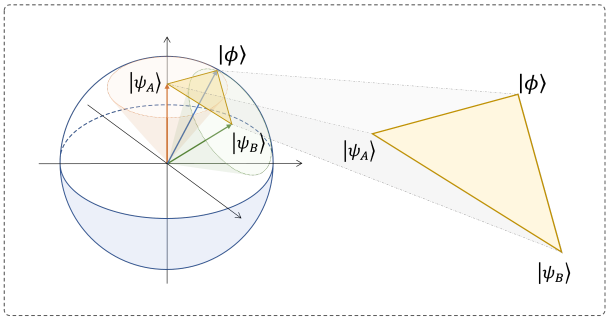

Figure 1: Schematic illustration of the geometric relationship between quantum states in Hilbert space revealed by the interference term.

The interference term is a unique feature derived from quantum probability, which reveals the geometric relationship of quantum states in Hilbert space, as shown in Fig. 1. α and β in the interference item are the parameters for constructing the superposition state, which can be regarded as weights. $|\langle\psi_A|\phi\rangle|$ and $|\langle\phi|\psi_B\rangle|$ describe the side lengths of the two sides of the triangle in the illustration, and $\cos(\theta)$ is the degree of an angle of the triangle. They jointly describe the area of the triangle, so the interference term itself can be considered as a description of the geometric relationship of the quantum state in Hilbert space.

In the learning model constructed in the above form, the interference term is used as a constraint item of the loss function to realize the spatial geometry constraint between the quantum states (that is, the representation obtained from the label data and the representations obtained from the non-homologous data). Compared with the traditional loss function with no constraint term, the loss function with the spatial geometry constraints can limit the value range of the features in the representation and alleviate the dependence on the large amount of data in the representation learning process. In fact, it is not difficult to understand that the constraint item describing the geometric relationship limits the positional relationship of the representations in space, which can naturally improve the difficulty of representation learning compared to unconstrained representations.

Methodology

Word Sense Disambiguation

The WSD task can be formalized as a mapping function from the word embedding of the target word in the disambiguated text, $V^{target} \in \mathbb{R}^{1\times n}$, to the text embedding of word sense definitions in the dictionary, $V_k^{definitions} \in \mathbb{R}^{1\times n}$,

$$\mathcal{F} : V^{target} \rightarrow V_k^{definitions} \tag{11}$$

where $V_k^{definitions}$ refers to the k-th word sense in the candidate list corresponding to the target word (Bevilacqua et al.

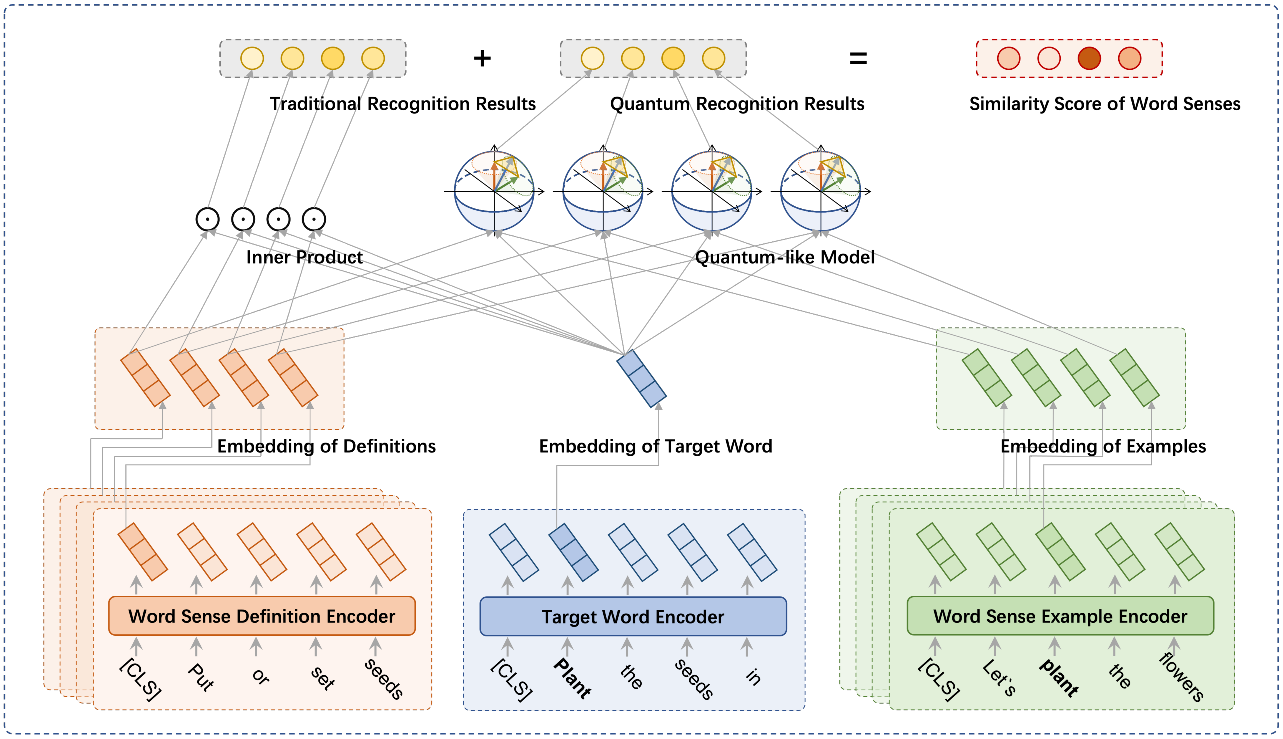

Figure 2: The architecture of QiWSD: a word sense disambiguation system with a traditional recognition method for high-frequency senses and a quantum recognition method for long-tail senses that can integrate non-homologous data using the quantum-like model.

2021; Navigli 2009; Zhang et al. 2024; Zhang, He, and Guo 2023).

Quantum-like Model with Quantum-inspired Constraint Mechanism

The core part of the quantum-like model with quantum-inspired non-homologous representation constraint mechanism is a quantum measurement operation, that is,

$$p(M_m; |\Psi\rangle) = \| M_m |\Psi\rangle \|^2, \tag{12}$$

in which the important components are the superposition state $|\Psi\rangle$ and the measurement operator M_m.

A superposition state can be composed of any number of quantum states $|\psi_i\rangle$,

$$|\Psi\rangle = \varepsilon_1|\psi_1\rangle + \varepsilon_2|\psi_2\rangle + ... + \varepsilon_i|\psi_i\rangle + ...; \tag{13}$$

the quantum states can be obtained by imposing a normalized function of the sum of squares on general representations V_i,

$$|\psi_i\rangle = SSN(V_i) = \frac{V_i}{\sqrt{\|V_i\|_2}}. \tag{14}$$

The representations used to construct the superposition state can be obtained from homologous or non-homologous data. However, the representations obtained based on homologous data do not have the spatial geometry constraints pointed out in this paper, because the representations obtained from homologous data will eventually be merged into one. It should be noted that the quantum-like model is also valuable for homogeneous data, which is equivalent to a model with data integration capabilities.

A measurement operator is constructed from a quantum state $|\phi\rangle$,

$$M_m = |\phi\rangle\langle\phi|; \tag{15}$$

the quantum state can also be obtained by applying $SSN(\cdot)$ to a general representation V_m,

$$|\phi\rangle = SSN(V_m). \tag{16}$$

QiWSD: Quantum-inspired WSD System

In this section, we apply the quantum-like model to build a WSD system to verify whether the spatial geometry constraint mechanism applicable to non-homologous data can improve the inaccurate representation of word sense definitions faced by long-tail WSD tasks. The quantum-inspired WSD system is called QiWSD, and its model structure is shown in Fig. 2.

BiWSD consists of two parts, the traditional recognition method for high-frequency senses and the quantum recognition method enhanced by non-homologous data for LTSs. The traditional recognition method has been verified to be effective for high-frequency WSD, and this part is added to make the overall WSD system take into account high-frequency senses. The quantum recognition method leverages the spatial geometry constraint mechanism proposed in this paper, and this part is added to make the overall WSD system take into account LTSs.

The traditional recognition method uses two pre-trained language models BERT (Devlin et al. 2019) as encoders (namely target word encoder and word sense definition encoder) to obtain the word embedding of the target word in the disambiguated text W^{text},

$$V^{target} = BERT_{Target}(W^{text}), \tag{17}$$

and the text embedding of the word sense definitions provided by the glosses $W^{glosses}$ in the dictionary,

$$V_k^{definitions} = BERT_{Glosses}(W_k^{glosses}). \tag{18}$$

According to the specification of the BERT model, the vector corresponding to the target word in the disambiguated text is used as the word embedding of the target word output by the target word encoder; the vector corresponding to the start token "[CLS]" in the gloss is used as the text embedding of the word sense definition output by the word sense definition encoder.

Finally, the inner product of the target word embedding V^{target} and word sense definition embeddings $V_k^{definitions}$ is calculated separately to obtain the similarity score of each word sense under the traditional recognition method,

$$Score_k^{Traditional} = V^{target} \odot V_k^{definitions}. \tag{19}$$

The quantum recognition method uses the target word embedding and the word sense definition embeddings respectively output by the target word encoder and word sense definition encoder. Furthermore, example sentences of word sense definitions from the dictionary are integrated for word sense definition embeddings using the quantum-inspired constraint mechanism proposed in this paper.

Similarly, a pre-trained language model BERT is used as a word sense example encoder to obtain the text embedding of example sentences $W^{examples}$,

$$V_k^{examples} = BERT_{Examples}(W_k^{examples}). \tag{20}$$

Since the target word exists in the example sentence, the vector corresponding to the target word in the example sentence is used as the text embedding output by the word sense example encoder.

Next, the quantum recognition method is implemented using the quantum-like model with the quantum-inspired constraint mechanism:

- $V_k^{definitions}$ and $V_k^{examples}$ are constructed as quantum states by the normalization function $SSN(\cdot)$,

  $$|\psi_k^{definitions}\rangle = SSN(V_k^{definitions}) \tag{21}$$

  and

  $$|\psi_k^{examples}\rangle = SSN(V_k^{examples}); \tag{22}$$

- they are constructed as a superposition state by Eq. (13),

  $$|\Psi_k^{def+exm}\rangle = \varepsilon_k^1|\psi_k^{definitions}\rangle + \varepsilon_k^2|\psi_k^{examples}\rangle \tag{23}$$

  where $\varepsilon_k^1 = \sin(\varepsilon_k)$, $\varepsilon_k^2 = \cos(\varepsilon_k)$, and $\varepsilon_k \in \mathbb{R}$ is obtained by $V_k^{definitions}$ and $V_k^{examples}$ through a linear layer of the neural network;

- V^{target} is constructed as a measurement operator by the normalization function $SSN(\cdot)$ and Eq. (15),

  $$|\phi^{target}\rangle = SSN(V^{target}) \tag{24}$$

  and

  $$M_+^{target} = |\phi^{target}\rangle\langle\phi^{target}| \tag{25}$$

  where "+" refers to the observations that the target word belongs to the corresponding word sense definition;

- finally, the similarity score of each word sense under the quantum recognition method is calculated through the quantum-like model Eq. (12),

  $$Score_k^{Quantum} = p(M_+^{target}; |\Psi_k^{def+exm}\rangle). \tag{26}$$

Model Training

We train QiWSD by optimizing the similarity scores of the word senses obtained by traditional and quantum recognition methods,

$$Score_k = a \cdot Score_k^{Traditional} + b \cdot Score_k^{Quantum}, \tag{27}$$

through cross-entropy loss,

$$\begin{aligned} &Loss(Score, index) \\ &= -\log\left(\frac{\exp(Score^{[index]})}{\sum_{i=1}\exp(Score^{[i]})}\right) \\ &= -Score^{[index]} + \log\sum_{i=1}\exp(Score^{[i]}), \end{aligned} \tag{28}$$

where $a, b \in \mathbb{R}$ are the weights of each recognition method, $Score = [Score_1, Score_2, ..., Score_k, ...]$, and $index$ is the index of the candidate list of the word senses.

Experiments

The following questions will be answered through experimental analysis:

- Whether the WSD system based on the quantum-like model can effectively improve the overall performance of the system is verified by standard and data-enhanced evaluation experiments;
- Whether the spatial geometry constraint mechanism for non-homologous data can effectively enhance the recognition ability of LTSs is verified by ablation experiments;
- Whether the system has effective generalization performance for other languages is verified by the latest cross-lingual datasets.

Datasets and Model Settings

Datasets: **The standard evaluation experiments** of QiWSD are constructed by using the WSD evaluation framework proposed by Navigli et al. (Navigli, Camacho-Collados, and Raganato 2017); **the data-enhanced evaluation experiments** are constructed by adding the training set WNGT[2]. The training set is SemCor[3], the development set is SemEval-07 (**SE7**; (Pradhan et al. 2007)), and the test

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

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

WSD Systems	Dev set	Test sets				Concatenation of all test sets				
	SE7	**SE2**	**SE3**	**SE13**	**SE15**	*Nouns*	*Verbs*	*Adj.*	*Adv.*	**ALL**
Standard Evaluation Experiments:										
EWISE (ACL; Kumar et al., (2019))	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
LMMS (ACL; Loureiro et al., (2019))	68.1	76.3	75.6	75.1	77.0	–	–	–	–	75.4
SREF (EMNLP; Wang et al., (2020))	72.1	78.6	76.6	78.0	80.5	80.6	66.5	82.6	84.4	77.8
ARES (EMNLP; Scarlini et al., (2020b))	71.0	78.0	77.1	77.3	**83.2**	80.6	68.3	80.5	83.5	77.9
BEM (ACL; Blevins et al., (2020))	74.5	79.4	77.4	79.7	81.7	81.4	68.5	**83.0**	**87.9**	79.0
EWISER (ACL; Bevilacqua et al., (2020))	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
SyntagRank (ACL; Scozzafava et al., (2020))	59.3	71.6	72.0	72.2	75.8	–	–	–	–	71.2
COF (EMNLP; Wang et al., (2021))	69.2	76.0	74.2	78.2	80.9	80.6	61.4	80.5	81.8	76.3
ESR (EMNLP; Song et al., (2021))	**75.4**	80.6	78.2	79.8	82.8	82.5	69.5	82.5	87.3	79.8
Z-Reweighting (ACL; Su et al., (2022))	71.9	79.6	76.5	78.9	82.5	–	–	–	–	78.6
Quantum-inspired Systems										
QWSD (RANLP; Tamburini, (2019))	–	70.5	69.8	69.8	73.4	73.6	54.4	77.0	80.6	70.6
DRWSD (CIKM; Zhang et al., (2022b))	74.7	80.8	78.0	80.0	82.7	82.7	69.5	82.9	86.6	80.4
QiWSD$_{\text{base}}$	74.8	81.0	79.3	80.8	82.7	83.7	71.5	82.8	87.6	80.8
QiWSD$_{\text{large}}$	75.2	**82.5**	**80.5**	**81.2**	**83.2**	**84.2**	**71.7**	**83.0**	87.7	**81.8**
Data-enhanced Evaluation Experiments:										
SparseLMMS (EMNLP; Berend, (2020))	73.0	79.6	77.3	79.4	81.3	–	–	–	–	78.8
EWISER (ACL; Bevilacqua et al., (2020))	75.2	80.8	79.0	80.7	81.8	81.7	66.3	81.2	85.8	80.1
ESR (EMNLP; Song et al., (2021))	**77.4**	81.4	78.0	**81.5**	**83.9**	83.1	71.1	**83.6**	87.5	80.7
QiWSD$_{\text{base}}$	75.0	81.8	79.5	81.0	83.0	84.0	72.2	82.8	87.6	81.1
QiWSD$_{\text{large}}$	75.4	**83.3**	**80.8**	**81.5**	**83.9**	**85.3**	**73.0**	83.0	**87.7**	**82.1**

Table 1: F1-score (%) on the English all-words WSD task. The comparison systems are divided into two groups: those under the standard evaluation experiments (i.e., using only SemCor) and those under the data-enhanced evaluation experiments (i.e., using SemCor and WNGT). SOTA performance is underlined compared to QiWSD$_{\text{base}}$ and bold compared to QiWSD$_{\text{large}}$.

sets include 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 the concatenation of all test sets (**ALL**).

Note that the candidate list of word senses includes all word sense definitions in WordNet 3.0. For the case where there are multiple word sense example sentences, the first one is selected by default; for the case where there is no word sense example sentence, the training text is used instead. Evaluation metrics and other unlisted information are subject to the settings of the evaluation framework.

Model Settings: The hardware platform deployed by QiWSD is Ubuntu 18.04, which installs six Tesla P40 GPUs. The development platform is Python 3.6, and the learning framework is Pytorch 1.8. WordNet 3.0 is provided by NLTK 3.5. Versions *bert-base-uncased* and *bert-large-uncased* of BERT are provided by Transformers 4.5. QiWSD based on *bert-base-uncased* and *bert-large-uncased* are called QiWSD$_{\text{base}}$ and QiWSD$_{\text{large}}$ respectively. *Learning rate*, *epoch* and *batch size* of QiWSD are $\{1e\text{-}5, 5e\text{-}6, 1e\text{-}6\}$, 20 and 4 respectively. Other hyperparameters not listed will be given in the published code.

Baselines

The comparison systems of **the standard evaluation experiments** select the related work of the past four years as the baselines, including EWISE (Kumar et al. 2019) and LMMS (Loureiro and Jorge 2019) in 2019, SREF (Wang and Wang 2020), ARES (Scarlini, Pasini, and Navigli 2020b), BEM (Blevins and Zettlemoyer 2020), EWISER (Bevilacqua and Navigli 2020) and SyntagRank (Scozzafava et al. 2020) in 2020, COF (Wang, Zhang, and Wang 2021) and ESR (Song et al. 2021) in 2021, Z-Reweighting (Su et al. 2022) in 2022. In addition, QWSD (Tamburini 2019) and DRWSD (Zhang et al. 2022b), which is also based on quantum theory, is selected. The comparison systems of **the data-enhanced evaluation experiments** include SparseLMMS (Berend 2020), EWISER (Bevilacqua and Navigli 2020) and ESR (Song et al. 2021). The experimental results of the above systems are all taken from the data published in the original papers.

Results and Analysis

The results of the standard and data-enhanced evaluation experiments are shown in Tab. 1.

From the perspective of overall performance, QiWSD outperforms the comparison systems in the standard evaluation experiments; QiWSD partially outperforms the comparison systems in the data-enhanced evaluation experiments. The reason for this phenomenon is that the training samples of LTSs are scarce under the standard experiment setting, but a certain number of training samples are provided under the data-enhanced experiment setting. It is conceivable that improving the weak position of LTSs by increasing the amount of data is directly effective for improving the recognition of LTSs. From the gap between the result values, QiWSD does not have a big gap with the comparison systems. The reason is that the number of LTSs is relatively small, and it is difficult to have a significant gap.

Models	Dev set	Test sets				
	SE7	SE2	SE3	SE13	SE15	ALL
Dataset: SemCor						
QiWSD$^{+}_{\text{base}}$	74.8	81.0	79.3	80.8	82.7	80.8
QiWSD$^{-}_{\text{base}}$	71.1	77.7	75.3	76.3	78.0	77.7
Dataset: LTS						
QiWSD$^{+}_{\text{base}}$	51.0	52.3	48.6	50.1	51.5	49.3
QiWSD$^{-}_{\text{base}}$	33.3	37.7	35.0	35.9	37.6	34.9

Table 2: Experimental results of the ablation experiments under the original (namely SemCor) and LTS datasets.

From the perspective of detailed performance, the result on the development set is suboptimal and the result on the test set **ALL** is optimal, indicating that QiWSD does not overfit the training data and has good generalization ability. The poor performance on the test sets *Adj.* and *Adv.* is due to the fact that there are relatively few LTSs in adjectives and adverbs, so QiWSD, which has the ability to recognize LTSs, cannot give full play to its advantages.

It should be emphasized that compared with the quantum-inspired systems QWSD and DRWSD, QiWSD is superior to the comparison systems in various indicators, indicating that it has certain competitiveness.

Ablation Study under LTS Datasets

Datasets: The datasets of the standard evaluation experiment setting and the constructed LTS datasets are used to carry out the ablation study. LTSs in the training set, development set and test sets of standard evaluation experiments are extracted and constructed as corresponding datasets. We refer to the word senses with less than three samples as LTSs.

Model Settings: Based on QiWSD$_{\text{base}}$, the model that deletes the component of the quantum recognition method is called QiWSD$^{-}_{\text{base}}$; the model that retains the component of the quantum recognition method is called QiWSD$^{+}_{\text{base}}$, which is the original model. Other information not listed remains the same as the above model settings.

Result Analysis: The experimental results are shown in Tab. 2, and the analysis is as follows:

- On the original datasets, the results of QiWSD$^{-}_{\text{base}}$ are obviously lower than those of QiWSD$^{+}_{\text{base}}$, which shows that the quantum recognition method is valuable and helpful to the overall performance of the WSD system. The reason for the gap of 3-4 percentage points in the result values is that the proportion of LTSs is relatively small.
- On the LTS datasets, the results of QiWSD$^{-}_{\text{base}}$ are also significantly lower than those of QiWSD$^{+}_{\text{base}}$, indicating that for the recognition of LTSs, the role of the quantum recognition method is significant. This is corroborated by a gap in the result values of around 15 percentage points.

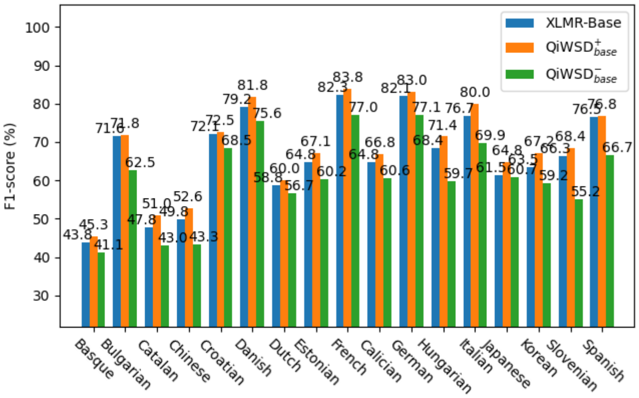

Figure 3: Experimental results of XLMR-Base (which results from data published by the evaluation framework), QiWSD$^{+}_{\text{base}}$ and QiWSD$^{-}_{\text{base}}$ under the cross-lingual datasets.

Experiments under Cross-Lingual Datasets

The generalization ability of QiWSD in other languages is verified under the latest cross-lingual datasets[4] proposed by Pasini et al. (Pasini, Raganato, and Navigli 2021). The performance in minority languages can also reflect the role of the quantum-like model from the side. The experimental models are QiWSD$^{+}_{\text{base}}$ and QiWSD$^{-}_{\text{base}}$ proposed by ablation experiments. The comparison model, XLMR-Base (Conneau et al. 2020), is the model used in the original paper. The encoders of the model are implemented by *bert-base-multilingual-cased* of BERT. Note that since the cross-lingual datasets are constructed based on BabelNet[5], the glosses and example sentences in this section are from BabelNet.

The experimental results are shown in Fig. 3. From the overall performance, QiWSD$^{+}_{\text{base}}$ is better than XLMR-Base to a certain extent and QiWSD$^{+}_{\text{base}}$ is definitely better than QiWSD$^{-}_{\text{base}}$, which shows that QiWSD has a certain generalization ability.

Conclusions

For long-tail WSD, the word sense definitions are limited, and it is difficult to obtain clear and easily distinguishable representations. Researchers propose to expand multi-source data to deal with it, but the unavoidable assumption is homogeneous data. This paper proposes a quantum-like model that simultaneously fuses representations obtained from homologous and non-homologous data, which means that the model has a stronger tolerance for multi-source data. The WSD system constructed based on the quantum-like model is verified under the WSD evaluation framework, the constructed LTS datasets and the cross-lingual datasets, and the experimental results show its effectiveness.

In future work, its internal mechanism and applicable fields will be further clarified. At the same time, systems based on the quantum-like model are constructed and verified on other tasks.

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

[5] https://babelnet.org/

Acknowledgements

Our work is supported by the National Key Research and Development Program of China (No. 2022YFC3600902) and the National Science Technology Major Project (No. 2018AAA0101904, 2023ZD0509706).

References

Abderrahim, M. A.; and Abderrahim, M. E. A. 2022. Arabic Word Sense Disambiguation for Information Retrieval. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 21(4): 69:1–69:19.

Basile, I.; and Tamburini, F. 2017. Towards quantum language models. In *EMNLP*, 1840–1849.

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.

Busemeyer, J. R.; and Bruza, P. D. 2012. *Quantum models of cognition and decision*. Cambridge University Press.

Campolungo, N.; Martelli, F.; Saina, F.; and Navigli, R. 2022. DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation. In *ACL*, 4331–4352.

Chen, Y.; Zhang, J.; and He, Q. 2022. PosWSD: Low-Resource Word Sense Disambiguation Model using Part Of Speech Information. In *IALP*, 26–31.

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. *ArXiv*, abs/1810.04805.

Du, Y.; Holla, N.; Zhen, X.; Snoek, C.; and Shutova, E. 2021. Meta-Learning with Variational Semantic Memory for Word Sense Disambiguation. In *ACL*, 5254–5268.

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

Farooq, U.; Dhamala, T. P.; Nongaillard, A.; Ouzrout, Y.; and Qadir, M. A. 2015. A word sense disambiguation method for feature level sentiment analysis. In *SKIMA*, 1–8.

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.

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

Kaddoura, S.; Ahmed, R. D.; and D., J. H. 2022. A comprehensive review on Arabic word sense disambiguation for natural language processing applications. *WIREs Data Mining Knowl. Discov.*, 12(4).

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, J.; Zhang, P.; Song, D.; and Hou, Y. 2016. An adaptive contextual quantum language model. *Physica A: Statistical Mechanics and its Applications*, 456: 51–67.

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.

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.

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.

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

Xie, M.; Hou, Y.; Zhang, P.; Li, J.; Li, W.; and Song, D. 2015. Modeling Quantum Entanglements in Quantum Language Models. In *IJCAI*, 1362–1368.

Yap, B. P.; Koh, A.; and Chng, E. S. 2020. Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences. In *EMNLP*, 41–46.

Zhang, G.; Lu, W.; Peng, X.; Wang, S.; Kan, B.; and Yu, R. 2022a. Word Sense Disambiguation with Knowledge-Enhanced and Local Self-Attention-based Extractive Sense Comprehension. In *COLING*, 4061–4070.

Zhang, J.; He, R.; and Guo, F. 2021. Bi-Matching Mechanism to Combat the Long Tail of Word Sense Disambiguation. In *ECMLPKDD*, 1–9.

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.; and Liu, C. 2024. Quantum Interference Model for Semantic Biases of Glosses in Word Sense Disambiguation. In *AAAI*, 19551–19559.

Zhang, J.; He, R.; Guo, F.; Ma, J.; and Xiao, M. 2022b. 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. 2021. 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.; Wang, J.; Wang, Y.; Hu, S.; Xiao, J.; and Li, Z. 2022c. Quantum Entanglement Inspired Correlation Learning for Classification. In *PAKDD*, volume 13281, 58–70.

Zhang, J.; Li, Z.; Xiao, J.; and Li, M. 2022d. Neural Network Model Reconstructed from Entangled Quantum States. In *IJCNN*, 1–8.