Bridging the Language Gap: Learning Distributed Semantics for Cross-Lingual Sentiment Classification

Guangyou Zhou[†], Tingting He[†] and Jun Zhao[‡]

†School of Computer, Central China Normal University,
152 Luoyu Road, Wuhan 430079, China
‡National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences

December 9, 2014

Outline

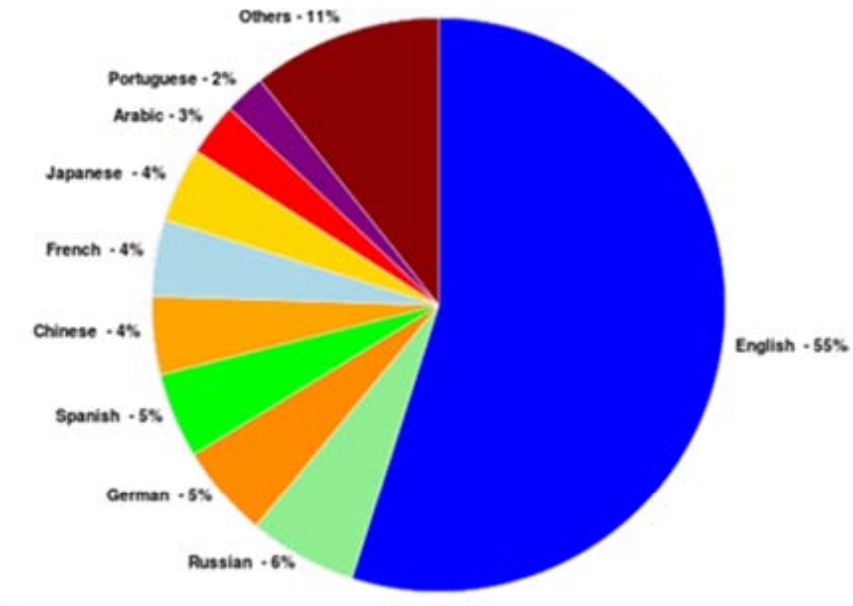
Multilingual Data

- More and more information becomes available in many different languages $\rightarrow$ multilingual information overload
- We need cheap and effective tools to search and organize these enormous amounts of multilingual information with minimum human input

Moving to Multilingual Settings

Why move to multilingual settings?

- Increased functionality (e.g., building translation resources, cross-lingual tasks)
- More available data in many different languages
- More reachable users



Moving to Multilingual Settings

Problems and challenges:

- High quality training datasets are more difficult to obtain
- However, most of them are English data
- Corpora for other languages, including Chinese, are rare

- We will focus on bilingual problems, adaptations to more languages are possible

- Cross-lingual Sentiment Classification

Problem Definition

Cross-lingual sentiment classification:

Learning how to label/categorize review data written in one language, and propagate sentiment labels to another language

- Knowledge transfer $\rightarrow$ learn/transfer sentiment labels from source to target
- Usually settings: source language is English; target language is Chinese
- Example use $\rightarrow$ Leverage Labeled English review data and Unlabeled Chinese review data

Problem Definition

Cross-lingual sentiment classification:

Learning how to label/categorize review data written in one language, and propagate sentiment labels to another language

- Knowledge transfer $\rightarrow$ learn/transfer sentiment labels from source to target
- Usually settings: source language is English; target language is Chinese
- Example use $\rightarrow$ Leverage Labeled English review data and Unlabeled Chinese review data

The fundamental challenge

A lack of overlap between the feature spaces of the source language data and that of the target language data.

Previous Solution

Pilot studies on cross-lingual sentiment classification:

Mihalcea et al. ACL 2007

- Bilingual lexicon and manually translated parallel corpus

Previous Solution

Pilot studies on cross-lingual sentiment classification:

Mihalcea et al. ACL 2007

- Bilingual lexicon and manually translated parallel corpus

Wan ACL 2009

- Borrowing statistical machine translation
- Training the model with the translated training data
- Labeling the training data with a co-training framework
- No much loss compared to human translation
- Suggesting MT is a viable way

Problems & Challenges

Machine translation-based approaches have certain issues:

- Machine translation may change the sentiment polarity
  - English sentence "it is too beautiful to be true" **Negative**
  - Translated Chinese: "实在是太漂亮是真实的" **Positive**
- Many sentiment indicative words cannot be learned from the translated labeled data
  - Due to the limited coverage of vocabulary in the machine translation results
  - Duh et al. (2011) reported a low overlap between the vocabulary of English documents and the documents translated from Japanese to English
- Translating all sentiment data from one language into the other languages is a time consuming and labor intensive job

Our Solution

We propose a deep learning approach (e.g., stacked autoencoders) to learn language-independent distributed representations:

- Our model is trained on a large-scale bilingual parallel data and then projects the source language and the target language into a bi-lingual space.
- The goal of our model is to learn distributed representations through a hierarchy of network architectures.
- The learned distributed representations can be used to bridge the gap between the source language and the target language.

Our Solution

Bilingual word embeddings

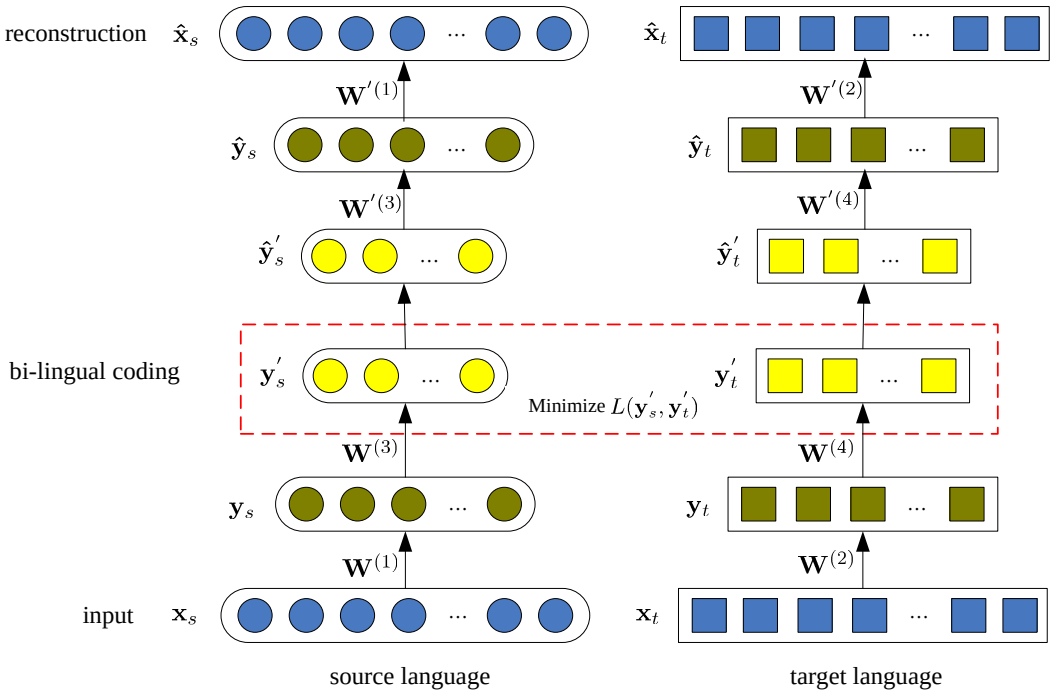
- Klementiev et al. (COLING 2012) leveraged bilingual word embeddings for cross-lingual sentiment classification.
- Zou et al. (ACL 2013) Leveraged bilingual word embeddings for phrase-based machine translation.
- A common property of these approaches is that a word-level alignment (e.g., GIZA++) of bilingual parallel corpus is used.

- Our approach only requires alignment parallel sentences and do not rely on word-level alignments during training, **which can simplifies the learning procedure and error propagation**.

Basic Idea

- Parallel data in multiple languages provides an alternative way for multiview representations, as parallel texts share their semantics.
- Given a large-scale parallel sentence pairs $(\mathbf{x}_s, \mathbf{x}_t)$, we would like to use it to learn distributed representations in both languages that are aligned.
- The idea is that a shared representation of two parallel sentences would be forced to capture the common information between two languages.

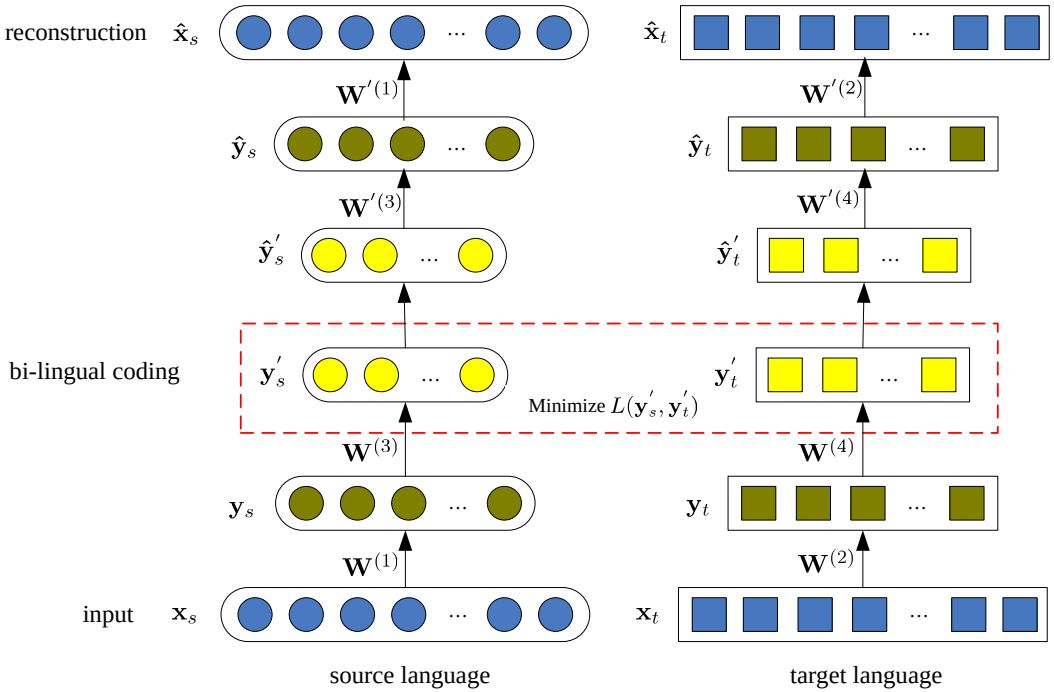
Our Framework



Input

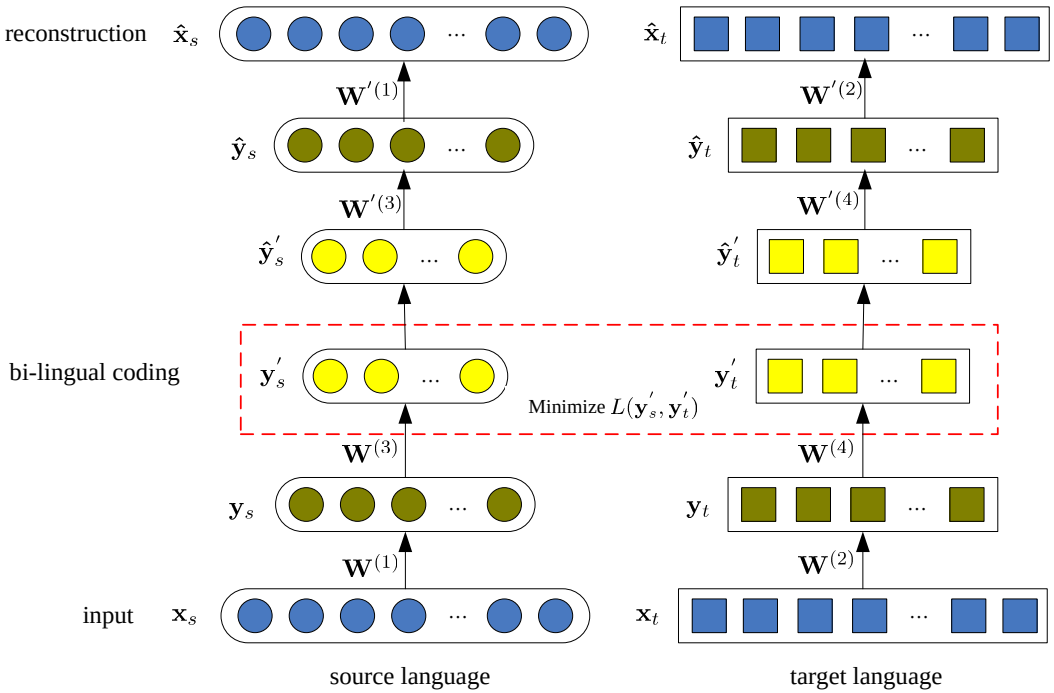
- A sentence with binary bag-of-words representation $\mathbf{x}_s$ in the source language.
- An associated binary bag-of-words representation $\mathbf{x}_t$ for the same sentence in the target language.

Encoding Phase



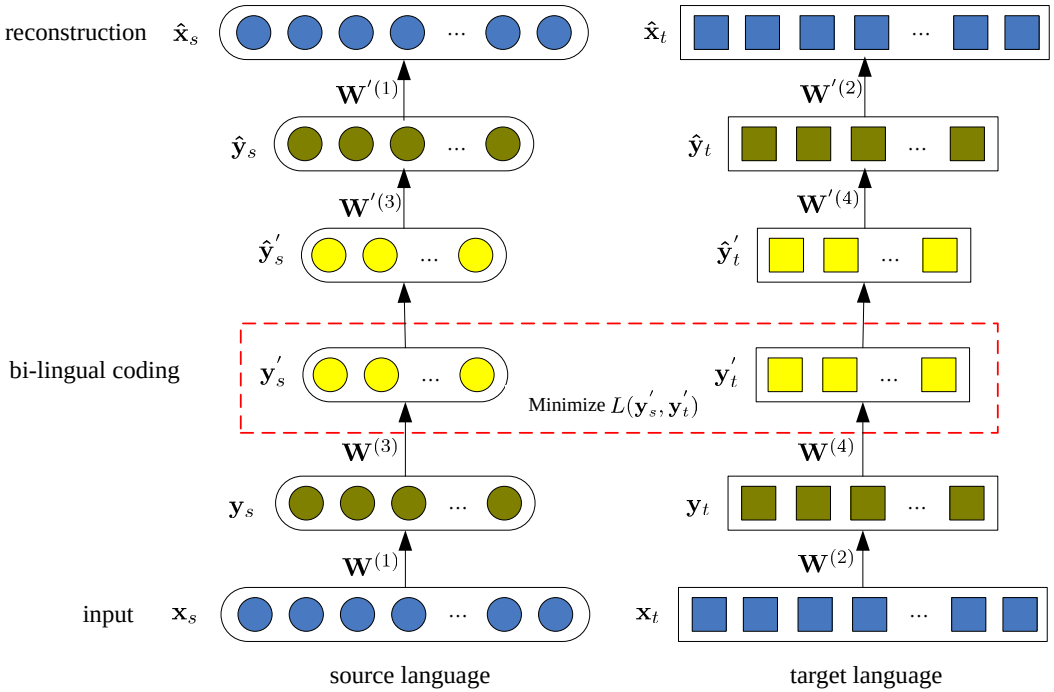
- We use the hyperbolic tangent function as the activation function for an encoder f_θ and a decoder $g_{\theta'}$.
- The high-level latent representations:
  $\mathbf{y}_s = f_{\theta_s}(\widetilde{\mathbf{x}}_s) = s(\mathbf{W}^{(1)}\widetilde{\mathbf{x}}_s + \mathbf{b}^{(1)})$,
  $\mathbf{y}'_s = f_{\theta_s}(\mathbf{y}_s) = s(\mathbf{W}^{(3)}\mathbf{y}_s + \mathbf{b}^{(3)})$.

Decoding Phase



- The decoding phase aims to perform a reconstruction of the original sentence in any of the languages.
- Given a decoder $g_{\theta'_s}$, we have $\hat{\mathbf{y}}'_\mathbf{s} = g_{\theta'_s} = s(\mathbf{W}'^{(5)}\hat{\mathbf{y}} + \mathbf{b}'^{(5)})$, $\hat{\mathbf{y}}_s = g_{\theta'_s} = s(\mathbf{W}'^{(3)}\hat{\mathbf{y}}' + \mathbf{b}'^{(3)})$, $\hat{\mathbf{x}}_s = g'_{\theta_s} = s(\mathbf{W}'\hat{\mathbf{y}}_s + \mathbf{b}'^{(1)})$.

Loss Function



- reconstruct $\mathbf{x}_s$ from itself (loss $L(\mathbf{x}_s, \hat{\mathbf{x}}_s)$).
- reconstruct $\mathbf{x}_t$ from itself (loss $L(\mathbf{x}_t, \hat{\mathbf{x}}_t)$).
- distance between the sentence level encoding of the bitext (loss $L(\mathbf{y}'_s, \mathbf{y}'_t)$).

Loss Function

The overall objective function is therefore the weight sum of these errors over a set of binary bag-of-words input vectors $\mathcal{C} = \{(\mathbf{x}_s^{(1)}, \mathbf{x}_t^{(1)}), \cdots, (\mathbf{x}_s^{(n)}, \mathbf{x}_t^{(n)})\}$:

$$J(\mathbf{x}_s, \mathbf{x}_t; \theta, \theta') = \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}} \left\{ L(\mathbf{x}_s, \hat{\mathbf{x}}_s) + L(\mathbf{x}_t, \hat{\mathbf{x}}_t) + L(\mathbf{y}'_s, \mathbf{y}'_t) + \frac{\lambda}{2}(\|\theta\|_2 + \|\theta'\|_2) \right. \tag{2.1}$$

- where L is a loss function, such as cross-entropy
- $\theta = \{\theta_s, \theta_t\}$ and $\theta' = \{\theta'_s, \theta'_t\}$ are the set of all model parameters
- we add the constraints $\mathbf{b}^{(1)} = \mathbf{b}^{(2)}$, $\mathbf{b}^{(3)} = \mathbf{b}^{(4)}$, $\mathbf{b}'^{(1)} = \mathbf{b}'^{(2)}$ and $\mathbf{b}'^{(3)} = \mathbf{b}'^{(4)}$ before the nonlinearity across encoders

Learning Algorithm

- Given $\theta = \{\theta_s, \theta_t\}$ and $\theta' = \{\theta'_s, \theta'_t\}$, the gradient becomes:

$$\frac{\partial L}{\partial \theta} = \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}} \frac{\partial J(\mathbf{x}_s, \mathbf{x}_t; \theta, \theta')}{\partial \theta} + \lambda\theta. \tag{2.2}$$

- Since the derivation of the minimization of the distance between the sentence-level bi-lingual coding of bitext and the reconstruction errors are not independent.
- L-BFGS can run over the unlabeled parallel data to minimize the objective function works well in practice.

Cross-Lingual Sentiment Classification

- Once we have learned the parameters θ and θ', we can transform the binary bag-of-words representation of the training data from the source language into the bi-lingual coding space.
- Then we train a simple sentiment classification model using a linear support vector machine (SVM) or other machine learning tools.
- For each of the test data from the target language, we also transform its bag-of-words representations into the bi-lingual coding space.
- Finally, we predict the sentiment polarity of the test data using the trained classification model.

Experimental Setup

Two cross-lingual sentiment classification settings

- No labeled data in the target language are available, we only use the labeled data in the source language.
- We have some labeled data in the target language, in order to make full use the labeled data in both languages.

Experimental Setup

Two cross-lingual sentiment classification settings

- No labeled data in the target language are available, we only use the labeled data in the source language.
- We have some labeled data in the target language, in order to make full use the labeled data in both languages.

Two cases

- One is English as the source language and Chinese as the target language.
- Another is Chinese as the source language and English as the target language.

Data Set

	MPQA	NTCIR-EN	NTCIR-CH
Positive	1,471 (30%)	528 (30%)	2,378 (55%)
Negative	3,487 (70%)	1,209 (70%)	1,916 (44%)
Total	4,958	1,737	4,294

Table: Statistics of data sets used in this paper.

Benchmark data sets (Lu et al., 2011; Meng et al., 2012)

- MPQA-EN (Labeled English Data)
- NTCIR-EN (Labeled English Data)
- NTCIR-CH (Labeled Chinese Data)

Data Set

	MPQA	NTCIR-EN	NTCIR-CH
Positive	1,471 (30%)	528 (30%)	2,378 (55%)
Negative	3,487 (70%)	1,209 (70%)	1,916 (44%)
Total	4,958	1,737	4,294

Table: Statistics of data sets used in this paper.

Four settings

- MPQA-EN $\rightarrow$ NTCIR-CH
- NTCIR-EN $\rightarrow$ NTCIR-CH
- NTCIR-CH $\rightarrow$ MPQA-EN
- NTCIR-CH $\rightarrow$ NTCIR-EN

Model Architecture

- To learn the parameters θ and θ', we use the Chinese-English parallel corpus (Munteanu and Marcu, 2005).
- The source language autoencoder and the target language autoencoder consist of 1000 hidden units.
- The second hidden layer with 500 latent units.
- The bi-lingual autoencoder containing 500 latent units.

Comparison Methods

- **SVM**: using monolingual labeled data.
- **MT-SVM**: borrowing statistical machine translation.
- **MT-Cotrain** (Wan, 2009 ACL).
- **Joint-Train** (Lu et al., 2011 ACL).
- **CLMM**: a generative cross-lingual mixture model (Meng et al., 2012 ACL).
- **DRW**: state-of-the-art method for cross-lingual sentiment classification (Klementiev et al., 2012 COLING).

Experimental Results: English → Chinese (No Chinese data is used)

Method	MPQA-EN → NTCIR-CH	NTCIR-EN → NTCIR-CH
SVM	N/A	N/A
MT-SVM	54.33	62.34
MT-Cotrain	59.11 (+4.78)	65.13 (+2.79)
Joint-Train	N/A	N/A
CLMM	71.52 (+17.19)	70.96 (+8.62)
DRW	72.27 (+17.94)	71.63 (+9.29)
DAEs	**72.85 (+18.52)**	**72.21 (+9.87)**

Table: Sentiment classification accuracy for Chinese only using English labeled data. Improvements of different methods over baseline MT-SVM are shown in parentheses.

Experimental Results: English → Chinese (Both labeled data are used)

Method	MPQA-EN → NTCIR-CH	NTCIR-EN → NTCIR-CH
SVM	80.58	80.58
MT-SVM	54.33 (-26.25)	62.34 (-18.24)
MT-Cotrain	80.93 (+0.35)	82.28 (+2.79)
Joint-Train	83.42 (+2.84)	83.11 (+2.53)
CLMM	83.02 (+2.44)	82.73 (+2.15)
DRW	83.54 (+2.96)	83.26 (+2.68)
DAEs	**83.81 (+3.23)**	**83.59 (+3.01)**

Table: Sentiment classification accuracy for Chinese by using English and Chinese labeled data. Improvements of different methods over baseline SVM are shown in parentheses.

Experimental Results: Chinese → English (No English data is used)

Method	NTCIR-CH → MPQA-EN	NTCIR-CH → NTCIR-EN
SVM	N/A	N/A
MT-SVM	52.47	58.51
MT-Cotrain	58.63 (+6.16)	63.72 (+5.21)
Joint-Train	N/A	N/A
CLMM	68.29 (+15.82)	69.15 (+10.64)
DRW	70.85 (+18.38)	72.57 (+14.06)
DAEs	**71.42 (+18.95)**	**73.38 (+14.87)**

Table: Sentiment classification accuracy for English only using Chinese labeled data. Improvements of different methods over baseline MT-SVM are shown in parentheses.

Conclusion I

- Distributed representations offer many opportunities for advanced **language-independent** and **language-pair-independent** document representations.

- Cross-lingual distributed representation learning:
  - a modeling concept convenient for large and unstructured cross-lingual (multilingual) document collections
  - sound theoretical foundation and interpretation
  - potential and utility in various (monolingual), cross-lingual and multilingual applications
  - may be trained on **non-parallel abundant data** $\rightarrow$ serve as useful knowledge sources for virtually any language pair (including community languages, unofficial languages)

Conclusion II

- The knowledge from the learned **cross-lingual distributed representation** is useful in many applications:
  - Cross-lingual information retrieval
  - Cross-lingual news clustering
  - Cross-lingual word sense disambiguation
  - Cross-lingual entity linking
  - Cross-lingual keyword and keyphrase extraction
  - ...
  - Cross-lingual **you-name-it task** (e.g., transliteration mining, summarization)

- **An open question** $\rightarrow$ Is it possible to apply the same modeling methodology in **multimodal settings**?

What to Take Home

- How to model heterogeneous data representation using different deep learning models
- How to model, train and infer cross-lingual or multilingual deep learning models
- What are latent cross-lingual codings
- How to bridge the gap between different languages using the latent cross-lingual codings
- How to use the models' output distributed representations in various cross-lingual or multilingual tasks
- How to use distributed knowledge in e-commerce applications

Challenges for You

- Building new or more powerful cross-lingual deep learning models by extending the basic models presented in this work
- Building more advanced classification models using distributed knowledge
- Applying cross-lingual deep learning models in new tasks and new domains

Thanks

- Once more, thanks to **General Chairs**, **PC Chairs**, **Area Chairs** and **the reviewers** for their hard work for our paper.

- This work is specially supported by **CCF Opening Projects of Chinese Information Processing**

- Finally, **THANK YOU** for attending! We hope that you've learned something today!