International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 29, No. 2 (2015) 1559003 (26 pages)

DOI: 10.1142/S021800141559003X

World Scientific
www.worldscientific.com

Document Representation with Statistical Word Senses in Cross-Lingual Document Clustering

Guoyu Tang* and Yunqing Xia†

Department of Computer Science and Technology
TNList, Tsinghua University
Beijing 100084, P. R. China
**sweetyuer@gmail.com*
†yqxia@tsinghua.edu.cn

Erik Cambria

School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798, Singapore
cambria@ntu.edu.sg

Peng Jin

School of Computer Science
Leshan Normal University
Leshan 614000, P. R. China
jandp@pku.edu.cn

Thomas Fang Zheng

Department of Computer Science and Technology
TNList, Tsinghua University
Beijing 100084, P. R. China
fzheng@tsinghua.edu.cn

Received 23 May 2014
Accepted 16 October 2014
Published 18 February 2015

Cross-lingual document clustering is the task of automatically organizing a large collection of multi-lingual documents into a few clusters, depending on their content or topic. It is well known that language barrier and translation ambiguity are two challenging issues for cross-lingual document representation. To this end, we propose to represent cross-lingual documents through statistical word senses, which are automatically discovered from a parallel corpus through a novel cross-lingual word sense induction model and a sense clustering method. In particular, the former consists in a sense-based vector space model and the latter leverages on a sense-based latent

† Corresponding author.

Dirichlet allocation. Evaluation on the benchmarking datasets shows that the proposed models outperform two state-of-the-art methods for cross-lingual document clustering.

Keywords: Word sense; cross-lingual document representation; cross-lingual document clustering.

1. Introduction

Economy globalization and internationalization of businesses urge organizations to handle an increasing number of documents written in different languages. As an important technology for cross-lingual information access, cross-lingual document clustering (CLDC) seeks to automatically organize a large collection of multi-lingual documents into a small number of clusters, each of which contains semantically similar cross-lingual documents.

Various document representation (DR) models have been proposed to deal with mono-lingual documents. The classic DR model is vector space model (VSM),[32] which typically makes use of words as feature space. However, words are in fact not independent of each other. Two semantic word relations are worth mentioning, i.e. synonymy and polysemy. Synonymy indicates that different words can carry almost all identical or similar meaning, and polysemy implies that a single word can have two or more senses. To address such issues, previous researches attempted to represent documents through either explicit or latent semantic spaces.[3,6,14,16,20,46]

In the cross-lingual case, DR models present two main issues: language barrier and translation ambiguity. As for the former, a term in one language and its counterparts in other languages should be viewed as a unique feature in cross-lingual DR. In some earlier systems, dictionaries were used to map cross-lingual terms.[12,25] However, such systems all suffered from the latter issue, which implies that one term can be possibly translated into different terms in another language, especially when such terms entail common-sense knowledge.[7] Two translation ambiguity scenarios are worth noting. In the first scenario, the term carries different meanings (namely, senses). For example, the word *arm* has two general meanings: (1) the part of body from shoulder to hand, and (2) a thing that is used for fighting. Accordingly, the word *arm* should be translated into 手臂 (shou3 bi4, arm) in a context relating to human body, but it should be translated as 装备 (zhuang1 bei4, arm) in a military context. The second scenario applies when we have to select one of the many possible translations to convey a specific meaning. For example, as a part of the human body, the word arm can be also translated into 胳膊 (ge1 bo2, arm), which is not quite the same as 手臂 (shou3 bi4, arm).

This is a common problem in natural language processing (NLP) research even in mono-lingual documents, e.g. when switching between different domains.[43] In the context of CLDC, popular approaches consist in exploring word co-occurrence statistics within parallel/comparable corpora.[18,23,35,45] Recent works improved clustering performance by aligning terms from different languages at topic-level.[4,27,29,41] Nonetheless, cross-lingual topic alignment still remains an open challenge.

In this work, we treat translation ambiguity of terms, e.g. 手臂 (shou3 bi4, arm) and 装备 (zhuang1 bei4, arm), as polysemy and translation choices, e.g. 手臂 (shou3 bi4, arm) and 胳膊 (ge1 bo2, arm), as synonyms. As synonymy and polysemy problems are closely related to word senses, we propose to represent document with cross-lingual statistical senses. Unlike previous approaches, which extract word senses from dictionaries, we propose to induce word senses statistically from corpora. To deal with cross-lingual cases, a novel cross-lingual word sense induction (WSI) model, referred to as CLHDP, is proposed to learn senses for each word (referred to as local word senses) respectively in parallel corpora. Thus, a sense clustering method is adopted to discover global word senses with semantic relatedness between senses of different words.

In this work, two cross-lingual DR models are proposed: a sense-based VSM and sense-based latent Dirichlet allocation (LDA) model. Two advantages of the proposed models are worth noting. Firstly, synonymy can be naturally addressed when word senses are involved. Words in one language that carry the same meaning can be organized by one word sense. In the cross-lingual case, words in cross languages can also be organized by one cross-lingual word sense. With synonymy addressed, cross-lingual documents can be more accurately represented. As a result, more accurate cross-lingual document similarity can be obtained and, hence, CLDC improves. Secondly, polysemy can also be well addressed as the translation ambiguity of polysemous words can be resolved within the cross-lingual contexts. Consequently, cross-lingual document similarity can be calculated more accurately when cross-lingual word disambiguation is achieved. By jointly addressing synonym and polysemy, the proposed cross-lingual DR models work at a more semantic-level and, thus, are able to outperform bag-of-words models.[8] Compared to topic-level DR models, moreover, the proposed models result to be more fine-grained and, hence, more accurate.

The structure of the paper is as follows: Section 2 introduces related work in the field of CLDC. Sections 3 and 4 illustrate in detail the proposed model. Section 5 presents evaluation and discussion. Section 6, finally, gives concluding remarks and future directions.

2. Related Work

2.1. *DR models*

This work is closely related to DR models. In traditional VSM, it is assumed that terms are independent from each other and, thus, any semantic relations between them are ignored. Previous works used concepts or word clusters[10,30] as features or used similarities of words,[13,42] but they still failed to handle the polysemy problem.

To address both of the synonymy and polysemy issues, some DR models are based on lexical ontologies such as WordNet or Wikipedia, to represent documents in a concept space.[14,16,17] However, the lexical ontologies are difficult to construct and are also hardly complete, moreover they tend to over-represent rare word senses, while missing corpus specific senses.

Representative extensions of the classic VSM are latent semantic analysis (LSA)[20] and LDA.[3] LSA seeks to decompose the term-document matrix by applying singular value decomposition, in which each feature is a linear combination of all words. However, LSA cannot solve the polysemy problem. LDA has successfully been used for the task of topic discovery[3,21] but, according to Ref. 24, it may not perform well by itself in text mining task, especially in the case of tasks requiring fine granularity discrimination, e.g. document clustering. Most of these semantic models, moreover, are designed for mono-lingual document sets, and cannot be used in cross-lingual scenarios directly.

2.2. *Cross-lingual document clustering*

The main issue of CLDC is dealing with the cross-language barrier. The straightforward solution is document translation. In TDT3, four systems attempted to use Machine Translation systems.[22] Results show that using a machine translation tool leads to around 50% performance loss, compared with mono-lingual topic tracking. This ascribed mainly to the poor accuracy of machine translation systems.

Dictionaries and corpora are two popular ways to get cross-language information. Some researches use dictionaries to translate documents.[12] Others use dictionaries to translate features or keywords. Mathieu *et al.* use bi-lingual dictionaries to translate named entities and keywords and modified the cosine similarity formula to calculate similarity between bi-lingual documents.[25] Pouliquen *et al.* rely on a multi-lingual thesaurus called Eurovoc to create cross-lingual article vectors.[31] However, it is hard to select proper translation of ambiguous words in different contexts.

To solve such a problem, some researches leverage on word co-occurrence frequencies from corpora.[12,25] However, they still need a dictionary but the human-defined lexical resources are difficult to construct and are also hardly complete. Wei *et al.* use LSA to construct a multi-lingual semantic space onto which words and document in either language can be mapped and dimensions are reduced again according to documents to be clustered.[41] Yogatama and Tanaka-Ishii use a propagation algorithm to merge multi-lingual spaces from comparable corpus and spectral method to cluster documents.[45] Li and Shawe-Taylor use Kernel Canonical Correlation Analysis, a method that finds the maximally correlated projections of documents in two languages for cross-language Japanese-English patent retrieval and document classification.[23]

Unlike document classification, document clustering usually lacks training data. Hence, semantic spaces are constructed from parallel/comparable corpora, and dimensions are selected on the basis of their importance in such corpora, which are usually different from the target multi-lingual documents. Mimno *et al.* introduce a poly-lingual topic model that discovers topics aligned across multiple languages.[27] However, topics generated from a parallel corpus may be not aligned well to the topics discovered from the target document. Tang *et al.* use cross-lingual word similarity, but ignores the translation ambiguity problem.[39]

In this work, we view language barrier and translation ambiguity as synonymy and polysemy problems and propose to use statistic word senses to represent documents in different languages. Our proposed model can concurrently deal with the problems of synonymy and polysemy and, hence, outperform the state-of-the-art CLDC methods.

2.3. *WSI and disambiguation*

Many approaches have been proposed to address the word sense disambiguation (WSD) task.[11,26,28] The use of word senses has been proved to enhance performances on many NLP tasks.[38] However, the use of word sense requires manually compiled large lexical resources such as WordNet.

In many other cases, word senses are learned from corpora in an unsupervised manner, known as WSI. Many WSI algorithms have been proposed in the literature.[9] The Bayesian model proposed in Ref. 5 uses an extended LDA model to induce word senses. It outperforms the state-of-the-art systems in SemEval-2007 evaluation[1] by using a hierarchical Dirichlet process (HDP)[40] to induce word senses. Unlike LDA, which requires a specified number of topics, HDP is able to infer such number automatically. Apidianaki uses a bi-lingual corpus and take translation equivalent clusters as word senses.[2] It assumes that word instances with the equivalent translation carry the same meaning, which is not always true as instances of the same word with different meanings may be translated as the same word in another language.

WSI algorithms have already been integrated in information retrieval.[34,37] However, to the best of our knowledge, the above-mentioned works only consider senses of query words, while in document clustering senses of every word in the documents should be identified.

In this paper, we propose to induce cross-lingual word senses from a parallel corpus by means of a novel Bayesian sense induction model, termed CLHDP, which is hereby also exploited for WSD.

3. CLDC System

3.1. *An overview*

Figure 1 presents the workflow of our sense-based CLDC system. Firstly, senses of individual words (referred to as local word senses) in each language are induced from the parallel corpus by means of a cross-lingual WSI (CL-WSI) algorithm. As a result, we obtain a set of local word senses, each of which is represented by distribution of cross-language words. Secondly, after grouping cross-language local word senses in one set, a clustering algorithm is used to partition such a set and, hence, to obtain a few word sense subsets, each of which contains some semantically similar

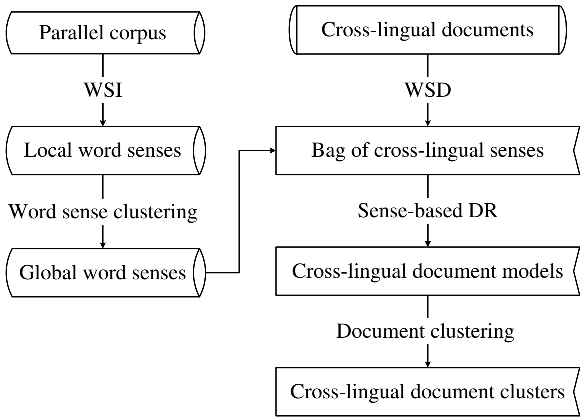


Fig. 1. Workflow of the CLDC system.

cross-language word senses. By using one sense for each subset to represent the different subsets, we obtain a few cross-lingual global word senses. Thirdly, cross-lingual documents are represented through such cross-lingual global word senses. Finally, the clustering algorithm is executed on the cross-lingual documents.

3.2. *Summary on novelty*

Two novel points in the CLDC system are worth noting.

(1) We propose a cross-lingual WSI algorithm by adapting mono-lingual HDP[40] to the cross-lingual scenario (CLHDP) and using a clustering method to discover semantic relatedness between senses of different words.
(2) Cross-lingual DR models are proposed to represent cross-lingual documents with the cross-lingual word senses, which are learnt by means of the CLHDP algorithm on a parallel corpus.

In the next sections, we show how and why the proposed models are better than existing DR models and explain in detail modules of the systems.

4. Theory and Algorithms

4.1. *Definitions*

Definition Local word sense
A local word sense s_w of word w is statistically represented by a set of discrete distributions over context words — one for a specific language l, i.e.:

$$s_w = \{c_i^l : p(c_i^l|s_w)\}; \quad i = 1, \ldots, N, \tag{1}$$

where s_w denotes a local sense of word w, c_i^l is a context word in language l, and $p(c_i^l|s_w)$ the probability distribution of c_i^l under s_w.

To obtain word senses, previous work relied on thesauri, which are time — and resourc consuming to construct. In this work, instead, we use context words, as well as their probabilities to reflect word senses. To use word *arm* again as an example, the following two local word senses can be learnt from the corpus.

- arm#1={limb: 0.159, forelimb: 0.069, sleeve: 0.019}
- arm#2={weapon: 0.116, war: 0.039, battle: 0.026}

The example indicates that a local sense of the word arm involves specific context words and their probability values, which are estimated from the corpus through a WSI algorithm. Obviously, local word senses can address the polysemy issue.

Definitions: Cross-lingual local word sense
In the cross-lingual scenario, we extend the local word sense definition so that it involves multi-lingual context words, which are extracted from a parallel corpus, i.e.:

$$s_w = \begin{bmatrix} \{c_i^{l_1} : p(c_i^{l_1}|s_w)\}, \quad i = 1, \dots, N_{l_1} \\ \cdots \\ \{c_j^{l_1} : p(c_i^{l_L}|s_w)\}, \quad j = 1, \dots, N_{l_L} \end{bmatrix}, \tag{2}$$

where $c_i^{l_k}$ is a context word in language l_k, and $p(c_i^{l_1}|s_w)$ the probability distribution of $c_i^{l_k}$ under s_w within texts in language l_k. For the word arm in the English-Chinese scenario, for example, the following two cross-lingual local word senses are illustrative.

- arm#1={limb: 0.159, forelimb: 0.069, sleeve: 0.019; 手臂: 0.137, 上肢: 0.079, 衣袖: 0.017}
- arm#2={weapon: 0.116, war: 0.039, battle: 0.026; 装备: 0.153, 武器: 0.027; 战争: 0.026}

With an English-Chinese parallel corpus, the cross-lingual local word senses can be obtained through the CL-WSI algorithm. As seen in the above example, the cross-lingual local word senses can address the polysemy issue in cross-lingual scenarios. However, local word senses are induced for every word separately. It is very common that a large number of synonymous word senses exist. Hence, we further propose to learn global word senses, which represent the universally exclusive word senses.

Definition: Cross-lingual global word sense
A global word sense g is a virtual word sense generalized from a group of synonymous local word senses, formalized as follows.

$$g = \{s_w^j\}; \quad j = 1, \dots, M, \tag{3}$$

where s_w^j represents a local word sense. When the local word senses are induced from a cross-lingual scenario, the global word sense becomes cross-lingual naturally. In our CLDC system, the global word senses are discovered through a clustering algorithm that uses context words as features in calculating semantic similarity between local word senses. Again, we use the word arm as an example to illustrate the global word sense:

- $g\#1$={*arm*#1, 手臂#1}={
  {limb: 0.159, forelimb: 0.069, sleeve: 0.019; 手臂: 0.137, 上肢: 0.079, 衣袖 : 0.017},
  {arm: 0.189, forelimb: 0.058, sleeve: 0.025; 胳膊: 0.159, 上肢: 0.089, 衣袖 : 0.014}
  }
- $g\#2$={arm#2, weapon#1, 装备#1}={
  {weapon: 0.116, war: 0.039, battle: 0.026; 装备: 0.153, 武器: 0.027; 战争 : 0.026},
  {arm: 0.12, battle: 0.04, war: 0.016; 装备: 0.133, 武器: 0.035; 战士: 0.028},
  {arm: 0.14, weapon: 0.12, war: 0.016; 装备: 0.133, 战争: 0.035; 战士: 0.028}
  }

As shown in the above examples, the senses arm#1 and 手臂#1 are organized by the global word sense $g\#1$ because the context distributions of arm#1 and 手臂 #1 are similar. In this way, synonymous word senses in both languages can be organized with one global word sense. Synonymy is thus successfully addressed. In the following sections, we present how the cross-lingual word senses are learned from the parallel corpus.

4.2. *Learning the cross-lingual word senses*

Two steps are required in learning the cross-lingual word senses:

(1) The local word senses are first induced from a parallel corpus;
(2) The global word senses are generalized from the local word senses.

4.2.1. *Local WSI*

The Bayesian model is adopted in order to achieve the task of local word sense learning. To be more specific, we extend HDP[40] to the cross-lingual scenario, referred to as CLHDP. Theory of HDP is briefly introduced first.

HDP for WSI

HDP is proposed to perform text modeling. Yao and Van Durme (2011) employ HDP for WSI.[44] HDP should be performed on each word respectively, which means each word has its own HDP model. In this paper, we define a word on which the WSI algorithm is performed as a target word. We also define words in the context of a target word as context words of the target word.

HDP is a generative model, which can randomly generate observed data. For each context v_i of the target word w, the sense s_{ij} for each word c_{ij} in v_i has a nonparametric prior G_i which is sampled from a base distribution G_w. H_w is a Dirichlet distribution with hyperparameter ϵ_w. The context word distribution η_{s_w} given a sense s_w is generated from H_w:$\eta_{s_w} \sim H_w$. The generative process of a target word w is given as follows:

(1) Choose $G_w \sim DP(\gamma_w, H_w)$.
(2) For each context window v_i of word w:
   (a) choose $G_i \sim DP(\rho_w, G_w)$.
   (b) for each context word c_{ij} of target word w:
      (i) choose $s_{ij} \sim G_i$.
      (ii) choose $c_{ij} \sim \mathrm{Mult}(\eta_{s_{ij}})$.

Hyperparameters γ_w and ρ_w are the concentration parameters for DP, controlling the variability of the distributions of G_w and G_i, respectively. HDP is illustrated in Fig. 2, where the shaded circle represents the observed variable, context word c_{ij}. HDP can be generated by the stick-breaking process and the Chinese restaurant process.[40]

CLHDP model

CLHDP models word senses through cross-lingual context tuples. Each tuple is a set of contexts that are equivalent to each other but written in different languages. Two assumptions are made in CLHDP. Firstly, contexts in a tuple share the same tuple-specific distribution over senses. Secondly, each sense consists of a set of discrete distributions over context words — one for each language $l = 1, \ldots, L$. In other words, rather than using a η_s for each sense s, as in HDP, there are L language-specific senses-context word distributions $\eta_s^1, \ldots, \eta_s^L$, each of which is drawn from a language-specific symmetric Dirichlet H_w^l with concentration parameter λ_w^l. CLHDP is illustrated in Fig. 3.

As shown in Fig. 3, the generative process of a target word w is given as follows:

(1) Choose $G_w \sim DP(\gamma_w, H)$.
(2) For each context window v_i of w:
   (a) choose $G_i \sim DP(\rho_w, G_w)$.
   (b) for each context word c_{ij}^l in language l of target word w:
      (i) choose $s_{ij}^l \sim G_i$.
      (ii) choose $c_{ij}^l \sim \mathrm{Mult}(\eta_{s_{ij}}^l)$.

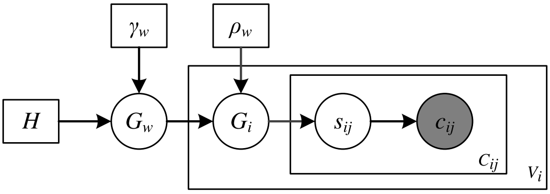


Fig. 2. Illustration of the CLHDP model.

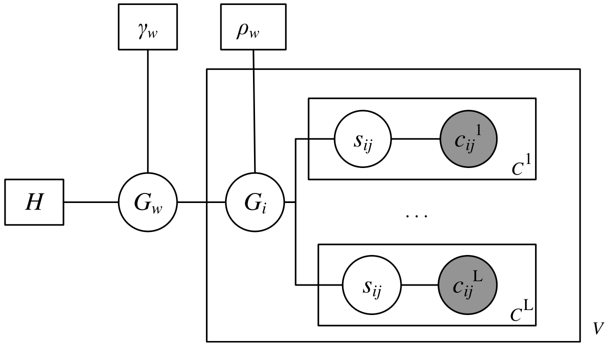


Fig. 3. Illustration of the CLHDP model.

Hyperparameters r_w and ρ_w are the concentration parameters of the DP, controlling the variability of the distributions G_w and G_{v_i}.

Inference for CLHDP model

Teh *et al.* use Collapse Gibbs Sampling to find latent variables in HDP. Gibbs Sampling initializes all hidden variable randomly.[40] For each iteration, hidden variables are sequentially sampled from the distribution conditioned on all other variables. Three sampling schemes can be used in HDP: posterior sampling in Chinese restaurant franchise, posterior sampling with an augmented representation and posterior sampling by direct assignment.

For CLHDP, we use the direct assignment scheme because it is easy to implement. There are three steps in sampling scheme:

(1) Given $\boldsymbol{s} = \{s_{ij}^l\}$ and $\boldsymbol{m} = \{m_{kj}\}$ in Chinese restaurant process, samples $\{G_v\}$ and G_w, where m_{kj} represents the number of tables in restaurant k serving dish j. The process is similar as described in Ref. 40.

The prior distribution G_w for each target word is a Dirichlet Process with concentration parameter λ_w and base probability H_w. It can be expressed using a stick-breaking representation,

$$G_w = \sum_{s_w=1}^{\infty} \pi_w^{s_w} \delta_{(\eta_{s_w}^1,\ldots,\eta_{s_w}^L)}, \tag{4}$$

where $\eta_{s_w}^1, \ldots, \eta_{s_w}^L$ are generated from $H_w^1, \ldots, H_w^L$ respectively and are given in this step, $\delta_{\eta_{s_w}^1,\ldots,\eta_{s_w}^L}$ is a probability measure concentrated at $\eta_{s_w}^1, \ldots, \eta_{s_w}^L$. $\{\pi_w^{s_w}\}$ are mixtures over senses. They are sampled from a stick-breaking construction. In the sampling process, suppose that we have seen S_w senses for the target word w. The context word distributions $\{\eta_{s_w}^l\}$ are generated and assigned to the context words in the corpus after some sampling iterations. G_w can be expressed as

$$G_w = \Sigma_{s_w} \pi_w^{s_w} \delta_{(\eta_{s_w}^1,\ldots,\eta_{s_w}^L)} + \pi_w^u G_w^u, \tag{5}$$

where G_w^u is distributed as Dirichlet Process $DP(\gamma_w, H_w)$. Thus, G_w is dependent on $\pi_{\boldsymbol{w}} = \{\pi_w^{s_w}\}$ and the sampling equation for $\pi_{\boldsymbol{w}}$ is as follows:

$$(\pi_w^1, \ldots, \pi_w^{S_w}, \pi_w^u) \mid \{\eta_{s_w}\}, \boldsymbol{s} \sim \mathrm{Dir}(m_{.1}, \ldots, m_{.S_w}, \gamma_w), \tag{6}$$

where $m_{.j}$ represents the number of tables in all restaurants serving dish j.

(2) Given $\{G_v\}$, G_w, sample $\boldsymbol{s} = \{s_{ij}^l\}$. The conditional probability for sampling the sense s_{ij} of context word $c_{ij}^l = c$ in context window v_i in language l can be estimated as:

$$P(s_{ij} = s_w, |\boldsymbol{s_{-ij}}, \boldsymbol{c_i}) = \begin{cases} (n_{-ij,s_w}^{v_i} + \rho_w \pi_w^{s_w}) \dfrac{n_{-ij,s_w}^c + \lambda_w^l}{n_{-ij,s_w,l} + V_{l,w}\lambda_w^l} & \text{if } s \text{ is previously used,} \\ \rho_w \pi_w^u \dfrac{n_{-ij,s_w}^c + \lambda_w^l}{n_{-ij,s_w,l} + V_{l,w}\lambda_w^l} & \text{if } s \text{ is new,} \end{cases} \tag{7}$$

where n_{-ij,s_w}^c is a count of how many context word $= c$ are assigned sense s_w, excluding the jth context word in language l and $V_{l,w}$ is the number of context words in language l. $n_{-ij,s_w,l}$ is the total number of context words in language l that are assigned sense s_w, excluding the jth context word in language l, $n_{-ij,s_w}^{v_i}$ is total number of context words in language l in v_i that are assigned sense s_w excluding the jth context word in language l.

(3) Given G_w, $\boldsymbol{s} = \{s_{ij}^l\}$, sample $\boldsymbol{m} = \{m_{kj}\}$. The conditional probability for sampling $\boldsymbol{m} = \{m_{kj}\}$ can be estimated as:

$$p(m_{kj} = m|\boldsymbol{s}, \boldsymbol{m_{-kj}}, \pi_{\boldsymbol{w}}) = \frac{\Gamma(\rho_w \pi_w^{s_w})}{\Gamma(\rho_w \pi_w^{s_w} + n_{k.j})} \mathrm{s}(n_{k.j}, m)(\rho_w \pi_w^{s_w})^m, \tag{8}$$

where $n_{k.j}$ represents the number of customers in restaurant k serving dish j, $\mathrm{s}(n_{k.j}, m)$ are unsigned Stirling numbers of the first kind.

Thus the context word distribution η_s^l can be calculated as

$$\eta_{s_w}^l(c) = \frac{n_{s_w,l}^c + \lambda_w^l}{n_{s_w,l} + V_{w,l}\lambda_w^l} \tag{9}$$

where $n_{s_w,l}^c$ is a count of how many context word $= c$ are assigned sense s, in language l and $V_{w,l}$ is the number of context words in language l. $n_{s_w,l}$ is the total number of words in language l that are assigned sense s_w.

In this work, we use sentences as context windows and extract cross-lingual context in a parallel corpus. For example, when a word is found in one sentence, we put the sentence and its corresponding sentence in the parallel corpus in a tuple.

4.2.2. *Global word sense generalization*

We view word sense generalization as a clustering task. The goal is to organize semantically similar word senses with one virtual word sense, which is globally unique.

In this work, probability distribution of context words is considered as a set of features and clustering algorithms are applied to merge equivalent senses. For a cross-lingual word sense, we simply combine context words in all languages and their distributions in one vector.

Two methods are adopted to cluster the local word senses:

(1) *Bisecting K-Means* is an extension of K-means, which is proven better than standard K-Means and hierarchical agglomerative clustering.[36] It begins with a large cluster consisting of every element to be clustered and iteratively picks the largest cluster in the set and splits it into two.
(2) *Graph-based Clustering* is a clustering method based on graph-partition. It first models the objects using a nearest-neighbor graph and then splits the graph into k-clusters using a min-cut graph-partitioning algorithm.

4.3. *Sense-based DR*

4.3.1. *Cross-lingual WSD*

In this work, the CLHDP algorithm is also used for WSD. Given $D = \{d_j; j = 1, \ldots, N\}$ representing a document set containing N documents and M words, the context set for each word is extracted and sense distribution in each context can be estimated by CLHDP model.

Given the word w in language l, the sense-context word distribution $\eta^l_{s_w}$ for word w estimated in parallel corpus, context sets $\{\hat{v}_i; i = 1, \ldots, \hat{V}_w\}$ in D, the inference process is similar as Sec. 4.2.1. The only modification is that in the second step, the conditional probability for sampling the sense s_{ij} of context word $c^l_{ij} = c$ in context window $\hat{v}_i$ in language l can be estimated as:

$$P(s_{ij} = s_w, |\boldsymbol{s_{-ij}}, \boldsymbol{c_i}) = (\hat{n}^{\hat{v}_i}_{-ij,s_w} + \hat{\rho}\hat{\pi}^{s_w}_{w})\eta^l_{s_w}(c_{ij}), \tag{10}$$

where $\hat{n}^{\hat{v}_i}_{-ij,s_w}$, $\hat{\rho}_w$, $\hat{\pi}^{s_w}_{w}$ represent CLHDP parameters in context sets $\{\hat{v}_i; i = 1, \ldots, \hat{V}_w\}$.

After sampling, the sense distribution $\theta_{\hat{v}}$ for each context window $\hat{v}_i$ in context set $\{\hat{v}_i; i = 1, \ldots, \hat{V}_w\}$ for the target word w can be estimated as follows:

$$\theta_{\hat{v}}(s_w) = \frac{\hat{n}^{\hat{v}}_{s_w} + \hat{\rho}_w\hat{\pi}^{s_w}_{w}}{\hat{n}^{\hat{v}} + \hat{\rho}_w \sum_{s'_w} \hat{\pi}^{s'_w}_{w}}, \tag{11}$$

where $\hat{n}^{\hat{v}}_{s_w}$ is a count of how many sense s_w in context window $\hat{v}$ and $\hat{n}^{\hat{v}}$ is the total number of words in context window $\hat{v}$.

With $\theta_{\hat{v}}(s_w)$, we simply take the mode sense in the distribution as the sense of the target word.

For example, three sentences are given below.

- S_1: *That man with one arm lost his other limb in an airplane crash.*
- S_2: *The nation must arm its soldiers for battle.*

- S_3: 国家必须为了战争武装它的士兵。

After stop word removal and word lemmatization, the three sentences become:

- $\overline{S}_1$: *man arm lost limb airplane crash*
- $\overline{S}_2$: *nation arm soldier battle*
- $\overline{S}_3$: 国家战争武装士兵

The probability of word sense arm#1 in sentence S_1 is 0.998005. For sentence S_2, The probability of word sense arm#2 is 0.944096.

In this work, we simply take the sense with the highest probability as the sense of the target word and use the senses to represent document. So the sense of arm in S_1 is $g\#1$ because the probability of arm#1 is higher and arm#1 belongs to $g\#1$. Similarly, the sense of arm in S_2 is $g\#2$.

For sentence S_3, the sense of 武装(wu3 qi4, arm) is also $g\#2$. In this way, instances of the same word with different meanings are identified as different senses and different words with same meaning are identified as the same sense. Accordingly, translation ambiguity and language barrier issues are both addressed.

After WSD, we start from the two most popular DR models, i.e. VSM and LDA, and propose sense-based versions of them.

4.3.2. *Sense-based VSM*

The traditional VSM model uses discriminative words to represent a document. Document d_i in document set D is represented as $d_i = \{w_{ij} : r_{ij}\}_{j=1,\ldots,M^{d_i}}$ in VSM, where r_{ij} represents the weight of a feature word w_{ij} in d_i. M^{d_i} is the number of feature words in d_i.

Differently, sense-based VSM (sVSM) uses global senses as features. With WSD, every word first in the document is assigned a unique global sense. Then, the weight of a global sense is calculated in a similar manner using TF-IDF formula. Finally, the sense vector is produced for each document. For example, d_i can be represented as $d_i = \{g_{ij} : \hat{r}_{ij}\}_{j=1,\ldots,M^{d_i}}$ in sVSM, where $\hat{r}_{ij}$ represents the weight of sense g_{ij} in d_i. We use cosine similarity to calculate similarities between two sense vectors.

4.3.3. *Sense-based LDA*

We replace word surfaces with word senses so that the classic LDA model is extended to sense-based LDA (sLDA) model. The WSD algorithm is again used to assign a unique global sense to a specific surface word. Then, sLDA generates a distribution of topics θ_{d_i} for each document d_i in the document set. For a word w_j in the document, the sense s_{ij} is drawn from the topic and topic-sense distribution ϕ containing T multinomial distributions over all possible senses in the corpus drawn from a symmetric Dirichlet distribution $\mathrm{Dir}(\beta)$.

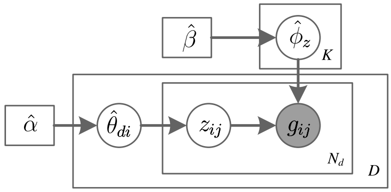


Fig. 4. Illustration of the sLDA model.

As shown in Fig. 4, the formal procedure of generative process in sLDA is given as follows:

(1) For each topic z:
  (a) choose $\hat{\phi}_z \sim \mathrm{Dir}(\hat{\beta})$.
(2) For each document d_i:
  (a) choose $\hat{\theta}_{d_i} \sim \mathrm{Dir}(\hat{\alpha})$.
  (b) for each word w_j in document d_i:
    (i) choose topic $z_{ij} \sim \mathrm{Mult}(\hat{\theta}_{d_i})$.
    (ii) choose sense $g_{ij} \sim \mathrm{Mult}(\hat{\phi}_{z_{ij}})$.

In sLDA, Gibbs sampling is used for parameter estimation and inference.[15] Compared with LDA, we replace the surface words with the induced word senses. Therefore, the topic inference is similar to the classic LDA, where the condition probability $P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{s})$ is evaluated by

$$P(z_{ij} = z|\boldsymbol{z_{-ij}}, \boldsymbol{g}) \propto \frac{n^{d_i}_{-ij,z} + \alpha}{n^{d_i}_{-ij} + Z\alpha} \times \frac{n^{g}_{-ij,z} + \beta}{n_{-ij,z} + G\beta}. \tag{12}$$

In Eq. (12), $n^{d_i}_{-ij,z}$ is the number of words that are assigned topic z in document d_i; $n^{d_i}_{-ij}$ is the total number of words in document d_i; $n^{g}_{-ij,z}$ is the number of senses with sense g that are assigned topic z; $n_{-ij,z}$ is the total number of words assigned topic z; G is the number of senses for the dataset. $-ij$ in all the above variables refers to excluding the count for the sense of the jth word. Further details are similar to the classic LDA.[15]

4.4. *Cross-lingual document clustering*

Document clustering becomes naturally feasible when the documents are represented by cross-lingual word senses. As the clustering algorithm is not the focus of this work, we simply adopt Bisecting K-Means to cluster document for sVSM. For sLDA, each topic in the test dataset is considered a cluster. After the parameters are estimated, documents are clustered into topics with the highest probabilities.

5. Evaluation

5.1. *Setup*

Development dataset

We randomly extract 1M parallel sentence pairs from LDC corpora (i.e. LDC2004E12, LDC2004T08, LDC2005T10, LDC2003E14, LDC2002E18 LDC2005T06, LDC2003E07 and LDC2004T07) as our development data to get word senses.

Test dataset

Four datasets are used in this paper.

(1) TDT4 datasets: Following Kong and Graff,[19] we use two datasets which are extracted from TDT4 evaluation dataset.
(2) CLTC datasets: Two datasets are extracted from the CLTC.[39]

Table 1 presents statistics of the four datasets.

In our experiments, we only extract nouns and verbs to induce word senses because words of other types make little contribution in document clustering. We use TreeTagger[33] to do lemmatization and POS tagging for English word, and use ICTCLAS[a] to segment Chinese words and assign POS tags to these words. Word information of the four test datasets is presented in Table 2.

Evaluation metrics

We adopt the evaluation metrics proposed by Steinbach *et al.*[36] The evaluation metrics are defined as the maximum score for each cluster. Let A_i correspond to the set of articles in a human-annotated cluster c_i. Let A_j correspond to the set of articles

Table 1. Statistics of topic and story in the four datasets. In each cell, number of topics is on the left and number of stories on the right.

Dataset	TDT41 (2002)	TDT42 (2003)	CLTC1	CLTC2
English	38/1270	33/617	20/200	20/600
Chinese	37/657	32/560	20/200	20/600
Common	40/1927	37/1177	20/200	20/1200

Table 2. Word statistics in the four datasets.

Dataset	TDT41 (2002)	TDT42 (2003)	CLTC1	CLTC2
English	2414	1887	1651	1862
Chinese	5457	3548	1437	2255
Common	7871	5435	3088	4117

[a] http://www.ictclas.org/ictclas_introduction.html.

in a system-generated cluster c_j. We consider each topic in the dataset as a cluster. The score for each cluster is based on the pairwised evaluation as follows:

$$
\begin{aligned}
p_{i,j} &= \frac{|A_i \cap A_j|}{|A_i|} \; p_i = \max_j\{p_{i,j}\},\\
r_{i,j} &= \frac{|A_i \cap A_j|}{|A_j|} \; r_i = \max_j\{r_{i,j}\},\\
f_{i,j} &= \frac{2 \cdot p_{i,j} \cdot r_{i,j}}{p_{i,j} + r_{i,j}} \; f_i = \max_j\{f_{i,j}\},
\end{aligned}
\tag{13}
$$

where $p_{i,j}$, $r_{i,j}$ and $f_{i,j}$ represent precision, recall and F-measure for the pair of clusters c_i and c_j, respectively. The general F-measure of a system is the micro-average of all the F-measures ($\{f_i\}$) for the system-generated clusters.

System parameters

The proposed approach involves great flexibility in modeling empirical data. This, however, entails that several parameters must be instantiated. More precisely, our model is regulated by the following four kinds of parameters:

(1) WSI parameters: We set $\gamma \sim \text{Gamma}(1, 0.1)$, $\rho \sim \text{Gamma}(0.01, 0.028)$ and $\lambda = 0.1$ for every word and both languages.
(2) sVSM parameters: We set number of clusters as number of topics in each dataset.
(3) sLDA parameters: We set $\alpha = 50/\#\text{topic}$, $\beta = 0.1$ which are usually used in LDA. The topic number is also set as cluster number in each dataset. In all experiments, we let the Gibbs sampler burn in for 2000 iterations and subsequently take samples 20 iterations apart for another 200 iterations.
(4) Number of global senses: We choose to conduct experiments to observe how they influence the document clustering performance.

5.2. *Experiment 1: Different word sense clustering methods*

In this experiment, we aim to study how different word sense clustering (WSC) methods influence the system performance. We implement two systems of different WSC methods.

(1) **Bisecting K-Means with sVSM (BK-sVSM):** The system uses Bisecting K-means to cluster local word sense. sVSM is used to represent documents. Cosine similarity measure is used to calculate document similarity and Bisecting K-means is used to cluster documents.
(2) **Graph-based Clustering with sVSM (GC-sVSM):** The system uses Graph-based clustering method to cluster local word sense. Other setups are the same as BK-sVSM.

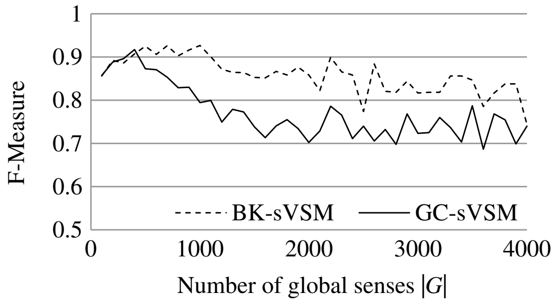


Fig. 5. Results of the systems with different sense numbers in CLTC1 test dataset.

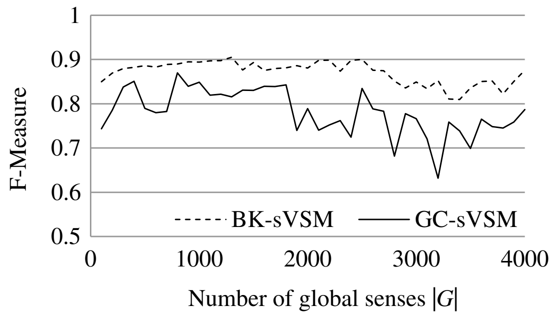


Fig. 6. Results of the systems with different sense numbers in CLTC2 test dataset.

We incrementally increase sense number $|G|$ from 100 to 4000 and evaluate both systems with the four datasets separately. Experimental results are presented in Figs. 5–8. The best F-measure values (f1) at the corresponding global word sense number (i.e. f1@Number #) of the two systems are listed in Table 3. The average efficiency of the two systems are listed in Table 4.

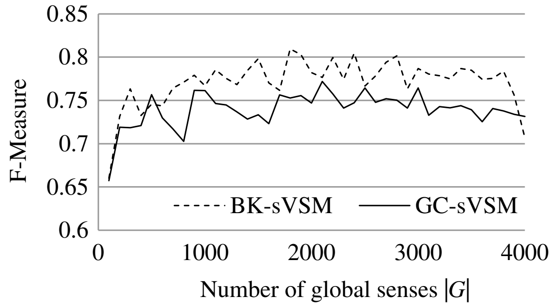


Fig. 7. Results of the systems with different sense numbers in TDT41 test dataset.

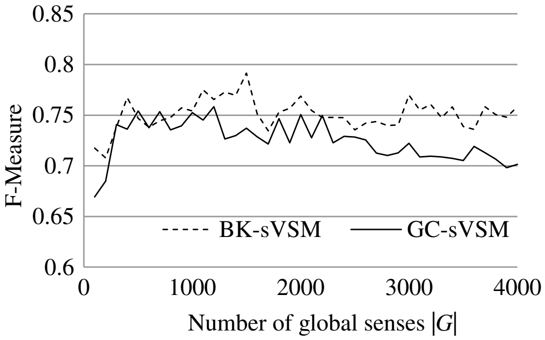


Fig. 8. Results of the systems with different sense numbers in TDT42 test dataset.

Table 3. The highest F-measure values of CLDC with different sense clustering methods.

System \ Dataset	CLTC1	CLTC2	TDT41	TDT42
BK-SVSM	**0.926@700**	**0.904@1300**	**0.809@1800**	**0.791@1500**
GC-SVSM	0.917@400	0.869@800	0.771@2100	0.752@1000

Table 4. The efficiency of CLDC with different sense clustering methods.

System \ Dataset	CLTC1	CLTC2	TDT41	TDT42
BK-SVSM	5753s	5732s	9623s	7129s
GC-SVSM	94s	144s	366s	222s

Discussion on influence of the global sense number

We compared the performance of different sense numbers and found that using low and high sense number can cause a drop on F-measure. This is due to the fact that, when the sense number is set to a low number, many local word senses that are not similar are clustered together resulting in low performance. When the sense number is set to a high number, similar local word senses are not clustered together. Thus, words with the same meaning in different languages may not be connected. This will largely affect the accuracy of similarity between documents in different languages. In that case, performance is reduced. After comparing different datasets, we can claim that datasets with larger word number have larger optimal global word sense number. For example, in system BK-sVSM, in CLTC1 dataset with 3088 words, the best F-measure achieves when the global word sense number is set as 700 while in TDT41 dataset with 7871 words, the optimal global word sense number is 1800. This is coherent with the fact that dataset with more words usually contains more senses.

Discussions on influence of the sense clustering method

As we can see from Figs. 5–8, BK-sVSM system outperforms GC-sVSM when the sense number $|G|$ increases from 100 to 3500 in most cases. This happens because the graph-based clustering method produces unbalanced clusters, while bisecting K-Means is more balanced in which it favors global property rather than the nearest neighbor. For this reason, we use Bisecting K-Means in WSC in the later experiments.

From Table 4 we can see GC-sVSM is much faster than BK-sVSM. This is because given n as the number of objects to be clustered, the time complexity of Bisecting K-Means is $\mathrm{O}(\mathrm{NNZ} * \log(k))$ where NNZ represent the number of nonzeros in the input matrix while the time complexity of Graph-based Clustering is $\mathrm{O}(n^2 + n * \mathrm{NNbrs} * \log(k))$ where NNbrs represents the number of neighbors in the nearest-neighbor graph. The number of target words is much smaller than the number of context words. So NNZ is much larger than n^2 and the time complexity of Bisecting K-Means is larger than the time complexity of Graph-based Clustering.

5.3. *Experiment 2: Different sense-based DR models*

In this experiment, we aim to study how different DR models influence system performance. Besides BK-sVSM, we also implement a system using sLDA to represent documents.

(1) **BK-sLDA:** The system uses sLDA to represent documents.

Experimental results on four datasets in two language cases are given in Figs. 9–12. The best F-measure (f1) at the corresponding global word sense number (i.e. f1@topic #) of the two systems are listed in Table 5.

Discussion

We compared the performance of different DR models based on sense and found that sVSM outperforms sLDA in all datasets. This is because fine granularity discrimination of feature space is important in document clustering task, while topic inferred from LDA may not resolve this issue very well. This is consistent with Ref. 24.

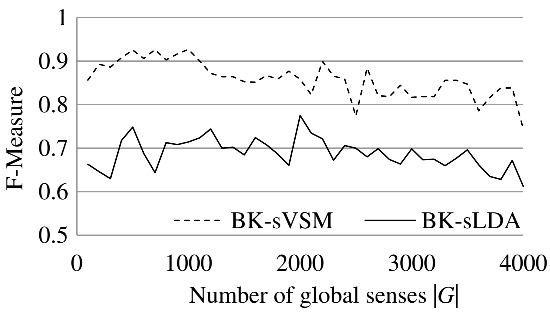


Fig. 9. Results of the systems with different sense numbers in CLTC1 test dataset.

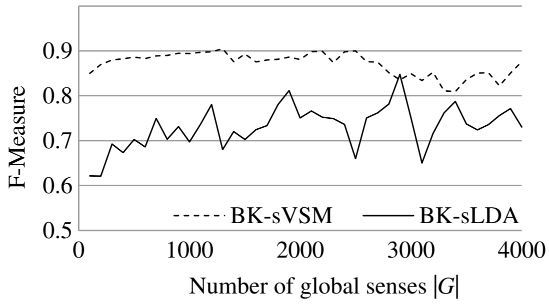


Fig. 10. Results of the systems with different sense numbers in CLTC2 test dataset.

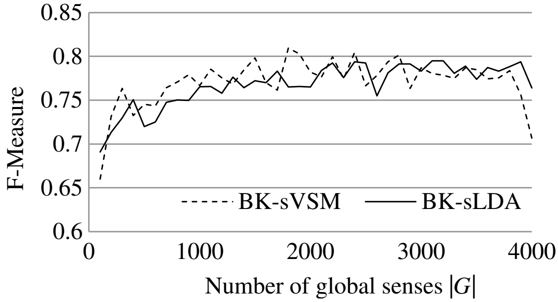


Fig. 11. Results of the systems with different sense numbers in TDT41 test dataset.

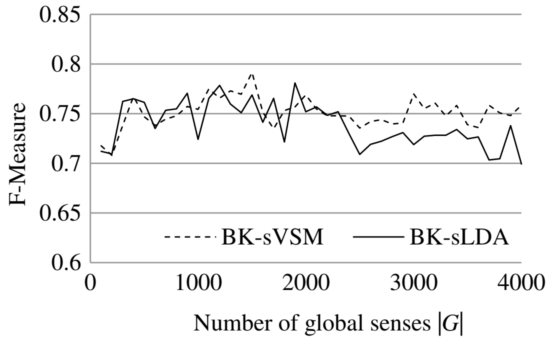


Fig. 12. Results of the systems with different sense numbers in TDT42 test dataset.

Table 5. The highest F-measure values of CLDC with different sense-based DR models in four test sets.

System \ Dataset	CLTC1	CLTC2	TDT41	TDT42
BK-SVSM	**0.926@700**	**0.904@1300**	**0.809@1800**	**0.791@1500**
GC-SVSM	0.774@2000	0.847@2900	0.795@3100	0.780@1900

5.4. *Experiment 3: Different document representation models*

In this experiment, we intend to compare our model with state-of-the-art DR models in CLDC. Besides BK-sVSM, the following two models are implemented.

(1) **CL-GVSM:** The model proposed by Tang *et al.* improves the similarity calculation by cross-lingual word similarity from a parallel corpus.[39]
(2) **PLTM:** The model proposed by Mimno *et al.* to get cross-lingual topic information.[27] In this paper, we apply the model on the parallel corpus to train cross-lingual topic and infer the topic distribution on the test dataset. The topic number is set to 1000. Bisecting K-means is used to cluster documents with the topic distribution as features.

Experiment results on four datasets in two language cases are given in Table 6.

Discussion

As shown in Table 6, BK-sVSM outperforms GVSM in all datasets. GVSM is a DR model that considers word similarity. However, it only considers relationships between words and ignores differences of one word in different contexts, which instead our proposed BK-sVSM considers both.

Table 6 also shows that BK-sVSM outperforms PLTM in all datasets. This indicates BK-sVSM outperforms PLTM in document clustering task. We find that PLTM yields the lowest performance in most cases. The reason for the significant performance drop is that, when using a parallel corpus to train the PLTM model, the topics may not be well covered in the test dataset and noise redundant topics (produced by the training corpus) may affect the performance.

Performance issue

Inducing word senses from the development data and clustering the word senses require higher computational effort. Indeed, the most time-consuming phase of our proposed model is the construction of word senses, which requires one CLHDP model for each word and a clustering method on those topics, referred to as local word senses in this paper. While word senses can be pre-computed or cached, word disambiguation of the test datasets still requires to be computed in real time. However, it can take advantage of parallel computing in which the disambiguation of each word is independent.

Table 6. The highest F-measure values with different DR models.

System \ Dataset	CLTC1	CLTC2	TDT41	TDT42
BK-SVSM	**0.926**	**0.904**	**0.809**	**0.791**
GVSM	0.900	0.898	0.762	0.748
PLTM	0.768	0.776	0.493	0.482

6. Conclusion

Previous researches show the importance of addressing language barrier and translation ambiguity in CLDC. In this paper, these two issues are viewed as general synonymy and polysemy problems and a DR based on cross-lingual statistical word senses is proposed.

The proposed method, in particular, aims to address the synonymy and polysemy issues in DR in two ways: (1) words containing the same meaning in different languages can be identified as the same word senses (in that case, language barrier can be crossed); (2) Instances of the same word with different meanings are identified as the different word senses (in that case, translation ambiguity can be addressed). Experiments on four datasets of two language cases show that our proposed model outperforms two state-of-the-art models in CLDC.

In the future, we plan to evaluate the performance of the proposed method with datasets of smaller samples. As the proposed method represents document in a word sense space, in fact, we can utilize it to handle sparse data problem with datasets of smaller samples, e.g. SMS messages and tweets.

Acknowledgment

This research work is supported by National Science Foundation of China (NSFC: 61272233). We thank the anonymous reviewers for their valuable comments.

References

1. E. Agirre and A. Soroa, Semeval-2007 task 02: Evaluating word sense induction and discrimination systems, in *Proc. 4th Int. Workshop on Semantic Evaluations* (Association for Computational Linguistics, 2007), pp. 7–12.
2. M. Apidianaki, Data-driven semantic analysis for multilingual wsd and lexical selection in translation, in *Proc. 12th Conf. European Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2009), pp. 77–85.
3. D. M. Blei, A. Y. Ng and M. I. Jordan, Latent dirichlet allocation, *J. Mach. Lear. Res.* **3** (2003) 993–1022.
4. J. Boyd-Graber and D. M. Blei, Multilingual topic models for unaligned text, in *Proc. Twenty-Fifth Con. Uncertainty in Artificial Intelligence* (AUAI Press, 2009), pp. 75–82.
5. S. Brody and M. Lapata, Bayesian word sense induction, in *Proc. 12th Conf. European Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2009), pp. 103–111.
6. E. Cambria, P. Gastaldo, F. Bisio and R. Zunino, An ELM-based model for affective analogical reasoning, *Neurocomputing* **149** (2015) 443–455.
7. E. Cambria, A. Hussain, C. Havasi and C. Eckl, Common sense computing: From the society of mind to digital intuition and beyond, in *Biometric ID Management and Multimodal Communication*, eds. J. Fierrez, J. Ortega, A. Esposito, A. Drygajlo and M. Faundez-Zanuy, Lecture Notes in Computer Science, Vol. 5707 (Springer, Berlin Heidelberg, 2009), pp. 252–259.
8. E. Cambria and B. White, Jumping NLP curves: A review of natural language processing research, *IEEE Comput. Intell. Mag.* **9**(2) (2014) 48–57.

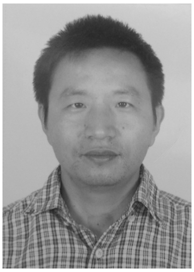
9. M. Denkowski, A survey of techniques for unsupervised word sense induction, *Language & Statistics II Literature Review* (2009) 1–8.
10. I. S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in *Proc. Seventh ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (ACM, 2001), pp. 269–274.
11. J. Duan, R. Lu and X. Li, Multi-engine collaborative boostrapping for word sense disambiguatoin, *Int. J. Artif. Intell. Tools* **16**(3) (2007) 465–482.
12. D. K. Evans and J. L. Klavans, A platform for multilingual news summarization, Technical Report (Department of Computer Science, Columbia University, 2003).
13. A. K. Farahat and M. S. Kamel, Statistical semantics for enhancing document clustering, *Knowl. Inf. Syst.* **28**(2) (2011) 365–393.
14. E. Gabrilovich and S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, in *Proc. IJCAI*, Vol. 7 (2007), pp. 1606–1611.
15. T. L. Griffiths and M. Steyvers, Finding scientific topics, *Proc. Nat. Acad. Sci. U. S. Am.* **101**(Suppl. 1) (2004) 5228–5235.
16. A. Hotho, S. Staab and G. Stumme, Ontologies improve text document clustering, in *Proc. ICDM 2003* (IEEE, 2003), pp. 541–544.
17. H.-H. Huang and Y.-H. Kuo, Cross-lingual document representation and semantic similarity measure: A fuzzy set and rough set based approach, *IEEE Trans. Fuzzy Syst.* **18**(6) (2010) 1098–1111.
18. K. Kishida, Double-pass clustering technique for multilingual document collections, *J. Inf. Sci.* **37**(3) (2011) 304–321.
19. J. Kong and D. Graff, Tdt4 multilingual broadcast news speech corpus, *Linguistic Data Consortium* (2005).
20. T. K. Landauer and S. T. Dumais, A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychol. Rev.* **104**(2) (1997) 211.
21. R. Y. K. Lau, Y. Xia and Y. Ye, A probabilistic generative model for mining cybercriminal networks from online social media, *IEEE Comput. Intell. Mag.* **9**(1) (2014) 31–43.
22. T. Leek, H. Jin, S. Sista and R. Schwartz, The BBN crosslingual topic detection and tracking system, *1999 TDT evaluation system summary papers* (Vienna, USA, 1999), pp. 214–221.
23. Y. Li and J. Shawe-Taylor, Advanced learning algorithms for cross-language patent retrieval and classification, *Inf. Process. Manag.* **43**(5) (2007) 1183–1199.
24. Y. Lu, Q. Mei and C. Zhai, Investigating task performance of probabilistic topic models: An empirical study of plsa and lda, *Inf. Retrieval* **14**(2) (2011) 178–203.
25. B. Mathieu, R. Besançon and C. Fluhr, Multilingual document clusters discovery, in *Proc. RIAO* (Citeseer, 2004), pp. 116–125.
26. R. F. Mihalcea and D. I. Moldovan, A highly accurate boostrapping algorithm for word sense diambiguation, *Int. J. Artif. Intell. Tools* **10** (2001) 5–21.
27. D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith and A. McCallum, Polylingual topic models, in *Proc. 2009 Conf. Empirical Methods in Natural Language Processing*, Vol. 2 (Association for Computational Linguistics, 2009), pp. 880–889.
28. R. Navigli and G. Crisafulli, Inducing word senses to improve web search result clustering, in *Proc. 2010 Conf. Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2010), pp. 116–126.
29. X. Ni, J.-T. Sun, J. Hu and Z. Chen, Mining multilingual topics from wikipedia, in *Proc. 18th Int. Conf. World Wide Web* (ACM, 2009), pp. 1155–1156.
30. J.-F. Pessiot, Y.-M. Kim, M. R. Amini and P. Gallinari, Improving document clustering in a learned concept space, *Inf. Process. Manag.* **46**(2) (2010) 180–192.

31. B. Pouliquen, R. Steinberger, C. Ignat, E. Kasper and I. Temnikova, Multilingual and cross-lingual news topic tracking, in *Proc. 20th Int. Conf. Computational Linguistics* (Association for Computational Linguistics, 2004), p. 959.
32. G. Salton, A. Wong and C.-S. Yang, A vector space model for automatic indexing, *Commun. ACM* **18**(11) (1975) 613–620.
33. H. Schmid, Probabilistic part-of-speech tagging using decision trees, in *Proc. Int. Conf. New Methods in Language Processing*, Manchester, UK, Vol. 12 (1994), pp. 44–49.
34. H. Schtze and J. O. Pedersen, Information retrieval based on word senses, *Proceedings for the Fourth Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, NV, 1995), pp. 161–175.
35. L. Shi, R. Mihalcea and M. Tian, Cross language text classification by model translation and semi-supervised learning, in *Proc. 2010 Conf. Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2010), pp. 1057–1067.
36. M. Steinbach, G. Karypis, V. Kumar *et al.*, A comparison of document clustering techniques, *KDD Workshop on Text Mining*, Boston, Vol. 400 (2000), pp. 525–526.
37. R. Steinberger, B. Pouliquen and C. Ignat, Newsexplorer: Multilingual news analysis with cross-lingual linking, in *Proc. of the 27th International Conference on Information Technology Interfaces* (2005).
38. C. Stokoe, M. P. Oakes and J. Tait, Word sense disambiguation in information retrieval revisited, in *Proc. 26th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (ACM, 2003), pp. 159–166.
39. G. Tang, Y. Xia, M. Zhang, H. Li and F. Zheng, Clgvsm: Adapting generalized vector space model to cross-lingual document clustering, in *Proc. IJCNLP* (2011), pp. 580–588.
40. Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei, Hierarchical dirichlet processes, *J. Am. Stat. Assoc.* **101** (2004) 1566–1581.
41. C.-P. Wei, C. C. Yang and C.-M. Lin, A latent semantic indexing-based approach to multilingual document clustering, *Decis. Support Syst.* **45**(3) (2008) 606–620.
42. S. K. M. Wong, W. Ziarko and P. C. N. Wong, Generalized vector spaces model in information retrieval, in *Proc. 8th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (ACM, 1985), pp. 18–25.
43. R. Xia, C. Zong, X. Hu and E. Cambria, Feature ensemble plus sample selection: A comprehensive approach to domain adaptation for sentiment classification, *IEEE Intell. Syst.* **28**(3) (2013) 10–18.
44. X. Yao and B. Van Durme, Nonparametric bayesian word sense induction, in *Proc. TextGraphs-6: Graph-based Methods for Natural Language Processing* (Association for Computational Linguistics, 2011), pp. 10–14.
45. D. Yogatama and K. Tanaka-Ishii, Multilingual spectral clustering using document similarity propagation, in *Proc. 2009 Conf. Empirical Methods in Natural Language Processing*, Vol. 2 (Association for Computational Linguistics, 2009), pp. 871–879.
46. L. Zhai, Z. Ding, Y. Jia and B. Zhou, A word position-related lda model, *Int. J. Pattern Recogn. Artif. Intell.* **25**(6) (2011) 909–925.

Guoyu Tang received her B.S. degree in Computer Science from the Department of Computer Science and Technology, Tsinghua University in 2009. She is currently a Ph.D. student in the Department of Computer Science and Technology, Tsinghua University. Her research interests include natural language processing, information retrieval and machine learning.

Yunqing Xia, IEEE senior member, received his Ph.D. in July 2001 from the Institute of Computing Technologies, Chinese Academy of Science. He worked as a postdoc research associate in the Natural Language Processing group at University of Sheffield, UK from January 2003 to October 2004. From December 2004 to September 2006, he worked as a postdoc fellow in the Department of Computer Science at the Chinese University of Hong Kong. In October 2006, he joined the Research Institute of Information Technology, Tsinghua University. His research interests include natural language processing, information retrieval and text mining. In the past 10 years, he has published more than 80 papers in distinguished journals (e.g. IEEE Intelligent Systems, IEEE Computational Intelligence) and conferences (e.g. ACL). He is also the inventor of three Chinese patents. He is currently serving as an editorial board member of Cognitive Computation Journal, advisory board member of Springer Socio-Affective Computing book series and associate editor of IJCPOL. He is co-organizer of ACM KDD WISDOM workshop and IEEE ICDM SENTIRE workshop.

Erik Cambria received his B.Eng. and M.Eng. with honours in Electronic Engineering from the University of Genoa in 2005 and 2008, respectively. In 2012, he was awarded his Ph.D. in Computing Science and Mathematics, following the completion of a Cooperative Awards in Science and Engineering (CASE) project born from the collaboration between the University of Stirling, the MIT Media Lab, and Sitekit Solutions Ltd., which included internships at HP Labs India, the Chinese Academy of Sciences, and Microsoft Research Asia. From August 2011 to May 2014, Erik was a research scientist at the National University of Singapore (Cognitive Science Programme) and an Associate Researcher at the Massachusetts Institute of Technology (Synthetic Intelligence Project). Today, Erik is an Assistant Professor at Nanyang Technological University (School of Computer Engineering), where he teaches natural language processing and data mining. He is also a research fellow at NTU Temasek Labs (TRF awardee), where he focuses on commonsense reasoning, concept-level sentiment analysis, and cyber-issue detection. Erik is an editorial board co-chair of Cognitive Computation, associate editor of Knowledge-Based Systems, and guest editor of many other top-tier AI journals, including three issues of IEEE Intelligent Systems, two issues of IEEE CIM, and one issue of Neural Networks. He is also involved in several international conferences as workshop series organizer, e.g. ICDM SENTIRE since 2011, program chair, e.g. ELM since 2012, PC member, e.g. UAI in 2014, tutorial organizer, e.g. WWW in 2014, special track chair, e.g. AAAI FLAIRS in 2015, and keynote speaker, e.g. CICLing in 2015.



Peng Jin received his B.S., M.S. and Ph.D. degrees in Computing Science from the Zhongyuan University of Technology, Nanjing University of Science and Technology, Peking University, respectively. From October 2007 to April 2008, he was a visiting student at the Department of Informatics, University of Sussex (funded by China Scholarship Council); from August 2014 to February 2015, he has been a visiting research fellow at the Department of Informatics, University of Sussex. Currently, he is an Associate Professor at Leshan Normal University (School of Computer Science). His research interests include natural language processing, information retrieval and machine learning.

Thomas Fang Zheng is a Full Research Professor and Vice Dean of the Research Institute of Information Technology, and the Director of the Center for Speech and Language Technologies, Tsinghua University. Since 1988, he has been working on speech and language processing, and has published over 230 journal and conference papers and 11 books. He holds 10 invention patents. He serves as Vice President — Institutional Relations and Education Program of APSIPA, and Director of the Speech Information Technical Commission of Chinese Information Processing Society of China. He is an IEEE Senior member, an Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing, an editor of Speech Communication, an Associate Editor of APSIPA Transactions on Signal and Information Processing, an Associate Editor of International Journal of Asian Language Processing, and a series editor of SpringerBriefs in Signal Processing. He also served as co-chair of Program Committee of International Symposium on Chinese Spoken Language Processing (ISCSLP) 2000, General Chair of Oriental COCOSDA 2003, Tutorial Co-Chair of APSIPA ASC 2009, General Co-Chair of APSIPA ASC 2011, APSIPA Distinguished Lecturer (2012–2013), General Co-Chair of IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) 2013, Area Chair of Interspeech 2014, and General Co-Chair of ISCSLP 2014.