Fine-grained Coordinated Cross-lingual Text Stream Alignment for Endless Language Knowledge Acquisition

Tao Ge[1,2], Qing Dou[3], Heng Ji[4], Lei Cui[2], Baobao Chang[1], Zhifang Sui[1], Furu Wei[2] and Ming Zhou[2]

[1]MOE Key Laboratory of Computational Linguistics, Peking University, Beijing, 100871, China
[2]Microsoft Research Asia, Beijing, 100080, China
[3]Facebook, CA, 94025, USA
[4]Rensselaer Polytechnic Institute, NY, 12180, USA
{tage,lecu,fuwei,mingzhou}@microsoft.com
douqing@gmail.com, jih@rpi.edu, {chbb,szf}@pku.edu.cn

Abstract

This paper proposes to study fine-grained coordinated cross-lingual text stream alignment through a novel information network decipherment paradigm. We use Burst Information Networks as media to represent text streams and present a simple yet effective network decipherment algorithm with diverse clues to decipher the networks for accurate text stream alignment. Experiments on Chinese-English news streams show our approach not only outperforms previous approaches on bilingual lexicon extraction from coordinated text streams but also can harvest high-quality alignments from large amounts of streaming data for endless language knowledge mining, which makes it promising to be a new paradigm for automatic language knowledge acquisition.

1 Introduction

Coordinated text streams (Wang et al., 2007) refer to the text streams that are topically related and indexed by the same set of time points. Previous studies (Wang et al., 2007; Hu et al., 2012) on coordinated text stream focus on discovering and aligning common topic patterns across languages. Despite their contributions to applications like cross-lingual information retrieval and topic analysis, such a coarse-grained topic-level alignment framework inevitably overlooks many useful fine-grained alignment knowledge. For example, Figure 1 shows typical knowledge that can be derived from fine-grained Chinese-English text stream alignments. In addition to (a) bi-lingual word translations, we can also discover (b) polysemous and multi-referential words if one Chinese word is aligned to multiple English words, (c) synonymous and co-referential word pairs if two Chinese words are aligned to the same English word, and (d) entity phrases (e.g., 阿布扎比 in Figure 1) if adjacent Chinese words in text are aligned to the same English named entity.

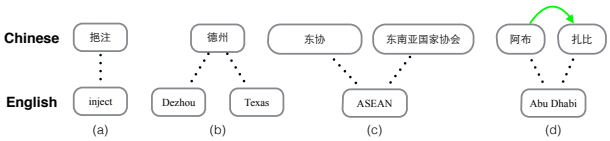

Figure 1: Knowledge derived from fine-grained cross-lingual text stream alignments: (a) word translations; (b) polysemy/multi-references; (c) synonym/co-reference; (d) entity phrases

In order to acquire language knowledge for Natural Language Processing (NLP) applications, we study fine-grained cross-lingual text stream alignment. Instead of directly turning massive, unstructured data streams into structured knowledge (D2K), we adopt a new Data-to-Network-to-Knowledge (D2N2K) paradigm, based on the following observations: (i) most information units are not independent, instead they are interconnected or interacting, forming massive networks; (ii) if information networks can be constructed across multiple languages, they may bring tremendous power to make knowledge mining algorithms more scalable and effective because we can employ the graph structures to acquire and propagate knowledge.

Based on the motivations, we employ a promising text stream representation – *Burst Information Networks (BINets)* (Ge et al., 2016a), which can be easily constructed without rich language resources, as media to display the most important information units and illustrate their connections in the text streams. With the BINet representation, we propose a simple yet effective network decipherment algorithm for aligning cross-lingual text streams, which can take advantage of the co-burst characteristic of cross-lingual text streams and easily incorporate prior knowledge and rich clues for fast and accurate network decipherment.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2496–2506
Brussels, Belgium, October 31 - November 4, 2018.

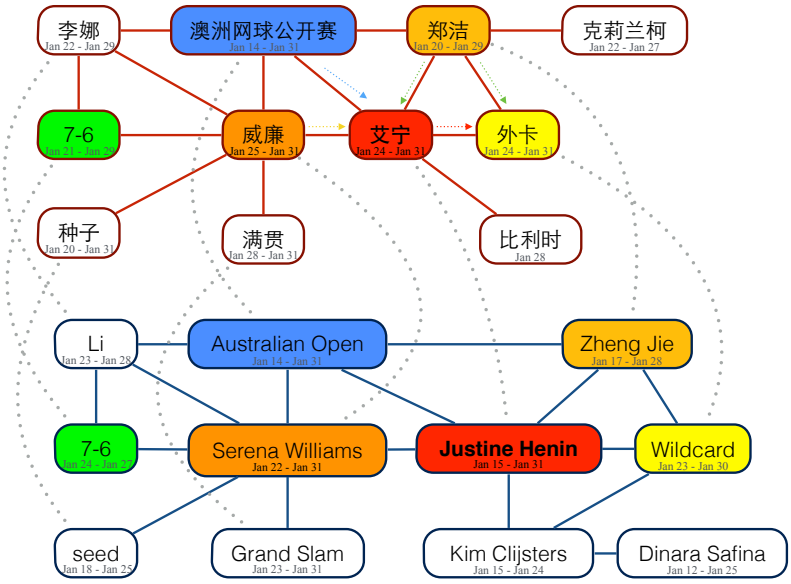

Figure 2: (A part of) Burst Information Networks built from Chinese and English news streams.

For example, in Figure 2, each node in a BINet is a bursty word with one of its burst periods, representing an important information unit in a text stream. To decipher the Chinese BINet, our approach first focuses on the nodes in the English BINet in Figure 2 as the candidates because they co-burst with the Chinese nodes. Then, we decipher some nodes based on prior knowledge (the green node), the pronunciation similarity clue (the orange nodes) or literal translation similarity clue (the blue node). These deciphered nodes will serve as neighbor clues to decipher their adjacent nodes (the red node) which will then be used for further decipherment (e.g., decipher the yellow node) through knowledge propagation across the network, as the dashed arrows in Figure 2 show.

Experiments on Chinese-English coordinated news streams show our approach can accurately align nodes across the cross-lingual BINets and derive various knowledge, and that with more streaming data provided, we can harvest more high-quality alignments and thus derive more knowledge. By aligning endless text streams, it is promising for never-ending language knowledge mining, which can not only complement language resources but also benefit some NLP applications.

The main contributions of this paper are:

- We propose a promising framework to mine knowledge from inexhaustible coordinated cross-lingual text streams through fine-grained alignment, exploring a paradigm for language knowledge acquisition.
- We propose a network decipherment approach for text stream alignment, which can work in both low and rich resource settings and outperform previous approaches.
- We release our data (annotations) and systems to guarantee the reproducibility and help future work improve on this task.

2 Burst Information Network

A Burst Information Network (BINet) is a graph-based text stream representation and has proven effective for multiple text stream mining tasks (Ge et al., 2016a,b,c). In contrast to many information networks (e.g., (Ji, 2009; Li et al., 2014)), BINets are specially for text streams. They focus on the burst information units which are usually related to important events or trending topics in text streams and illustrate their connections.

A BINet is originally defined as $G = \langle V, E, \boldsymbol{\omega} \rangle$ in (Ge et al., 2016a). Each node $v \in V$ is a burst element defined as a burst word[1] during one of its burst periods $\langle w, \mathcal{P} \rangle$ where w denotes a word and $\mathcal{P}$ denotes one consecutive burst period of w, as Figure 2 shows. Each edge $\epsilon \in E$ indicates the connection between two burst elements with the weight ω which is defined as the number of documents where these two burst elements co-occur in the text stream. In this paper, we extend the BINet definition to $G = \langle V, E, \boldsymbol{\omega}, \boldsymbol{\pi} \rangle$ by adding a binary

[1] Burst words and their corresponding burst periods can be detected based on Kleinberg burst detection algorithm (Kleinberg, 2003), as (Ge et al., 2016a) did.

indicator π to indicate if two nodes (i.e., burst elements) are frequently (more than 5 times) adjacent (as a bigram) in text, for mining knowledge such as entity phrases in Figure 1(d).

3 Decipherment

After constructing a BINet from a foreign language (we use Chinese as a foreign language in this paper), we can decipher it by consulting an Engish BINet constructed from its coordinated English text stream. We define $G_c = \langle V_c, E_c, \boldsymbol{\omega_c}, \boldsymbol{\pi_c} \rangle$ and $G_e = \langle V_e, E_e, \boldsymbol{\omega_e}, \boldsymbol{\pi_e} \rangle$ as the Chinese BINet and English BINet respectively. For people who do not know Chinese, G_c is a network of ciphers. We design a novel BINet decipherment procedure to decipher G_c by aligning as many nodes in G_c as possible to G_e. The decipherment process is defined to find $e \in V_e$ for a node $c \in V_c$ so that e is c's counterpart in the English text stream.[2]

3.1 Starting Point

To decipher the Chinese BINet, we need a few seeds based on prior knowledge as a starting point. Inspired by previous work on bi-lingual lexicon induction, decipherment and name translation mining, we utilize a few linguistic resources - a bi-lingual lexicon and language-universal representations such as time/calendar date, number, website URL, currency and emoticons to decipher a subset of Chinese nodes. For the example shown in Figure 2, we can decipher some nodes in the Chinese BINet such as "7-6" (to "7-6") and "种子" (to "seed").

3.2 Candidate Generation

For the nodes that cannot be deciphered by the prior knowledge, we first need to discover their possible candidates. For a node c in the Chinese BINet, its counterpart e can be any node in the English BINet or does not exist in the English BINet, resulting in an extremely large search space. Fortunately, burst information that refers to a hot topic usually co-bursts across languages. Based on this characteristic, for a node in the Chinese BINet, its counterpart is likely to be a node with the same burst period in the English BINet. For example, the node "威廉(Williams)" in the Chinese BINet in Figure 2 bursts between January 25 and January 31, 2010. We only need to look for its counterpart from the nodes in the English BINet whose burst period overlaps with this period. Formally, for a node $c \in V_c$ in the Chinese BINet, its candidate nodes in the English BINet can be derived as:

$$Cand(c) = \{e | \mathcal{P}(e) \cap \mathcal{P}(c) \neq \emptyset\}$$

where $e \in V_e$, and $\mathcal{P}(c)$ and $\mathcal{P}(e)$ are the burst periods of c and e respectively.

3.3 Candidate Verification

For the candidate list for c (i.e., $Cand(c)$), we need to verify each node $e \in Cand(c)$ and choose the most probable one as c's counterpart. Formally, we define $Score(c, e)$ as the credibility score of e being the correct counterpart of c and propose the following novel clues for verification.

Pronunciation

Inspired by previous work on name translation mining (e.g., (Schafer III, 2006; Sproat et al., 2006; Ji, 2009)), for a node $e \in Cand(c)$, if its pronunciation is similar to c, then e is likely to be the translation of c. For a Chinese node c and an English node e, we define S_p as its scaled pronunciation score to measure their pronunciation similarity whose range is $[0, 1]$:

$$S_p \in [0, 1] \propto \frac{1}{LD}$$

where LD is the normalized (by e's length) Levenshtein edit distance between c's pinyin[3] string and e's word string.

Translation

For a node $e \in Cand(c)$, it is possible that e's word exists or partially exists in the bi-lingual lexicon. We can exploit the translation clue to verify if e is c's counterpart. For example, "Australian Open" is a candidate of "澳洲网球公开赛(Australian Open)" as shown in Figure 2. Even though "澳洲网球公开赛(Australian Open)" is not in the bi-lingual lexicon, "Australian" and "open" are in the lexicon and their Chinese translations are "澳洲的(Australian)" and "公开(open)" respectively. If we literally translate "Australian Open" word by word, we will get "澳洲的公开" which has long common subsequences with the Chinese node "澳洲网球公开赛(Australian Open)", inferring that "Australian Open" is likely to be the translation of "澳洲网球公开赛".

[2] c and e are burst elements (i.e., nodes in the BINets). Sometimes, we also use c and e to denote the nodes' word if that does not lead to misunderstanding.

[3] Pinyin is the official romanization system for Chinese. We use pinyin instead of IPA because romanization is usually more easily available than IPA for a language.

Motivated by this observation, for a candidate $e \in Cand(c)$, we first extract its possible Chinese translations $C(e)$ from the bilingual lexicon. Note that if e is a multi-word, we concatenate translations of its components. Then, for $\langle c, e \rangle$, we define S_t as its scaled translation similarity score whose range is $[0, 1]$:

$$S_t \in [0, 1] \propto \max_{c' \in C(e)} LCS(c, c')$$

where $\max_{c' \in C(e)} LCS(c, c')$ is maximum length of the longest common subsequence between c and $c' \in C(e)$.

Neighbor

The graph topological structure of a BINet is also an important clue for decipherment. By analyzing a node's neighbors, we can learn useful topic-level knowledge to decipher the node. For the example in Figure 2, "艾宁(Henin)" in the Chinese BINet has neighbors such as "威廉(Williams)", "澳洲网球公开赛(Australian Open)" and "郑洁(Zheng Jie)" while "Justine Henin" in the English BINet is connected with "Serena Williams", "Australian Open" and "Zheng Jie". If we know "Serena Williams", "Australian Open" and "Zheng Jie" are the counterpart of '威廉", "澳洲网球公开赛" and "郑洁" respectively, we can infer "Justine Henin" is likely to be the counterpart of "艾宁", which can be further used as a clue to decipher its neighbors such as "外卡(wildcard)" through knowledge propagation.

We define $N(c)$ and $N(e)$ as the set of adjacent nodes of c in the Chinese BINet and the adjacent nodes of e in the English BINet respectively. The neighbor clue score S_n of $\langle c, e \rangle$ is defined as:

$$S_n = \sum_{c' \in N(c)} \hat{\omega}_{c,c'} \max_{e' \in N(e)} Score(c', e') \quad (1)$$

where $Score(c', e')$ is the overall score of e' being the counterpart of c', as defined at the beginning of this section, $\hat{\omega}_{c,c'} = \frac{\omega_{c,c'}}{\sum_{c'' \in N(c')} \omega_{c',c''}}$ is the normalized weight of the edge between c and c'.

Correlation of burst

If the word of $e \in Cand(c)$ frequently co-bursts with the word of c, then e is likely to be the counterpart of c. For example, "Serena Williams" in the English stream usually co-bursts with "小威" in the Chinese stream, as shown in Figure 3, which is a useful clue to infer that "Serena Williams" is the counterpart of "小威".

We define S_b as the burst correlation score:

$$S_b = \frac{\boldsymbol{s}_{w(c)} \cdot \boldsymbol{s}_{w(e)}}{\|\boldsymbol{s}_{w(c)}\|_1 + \|\boldsymbol{s}_{w(e)}\|_1 - \boldsymbol{s}_{w(c)} \cdot \boldsymbol{s}_{w(e)}} \quad (2)$$

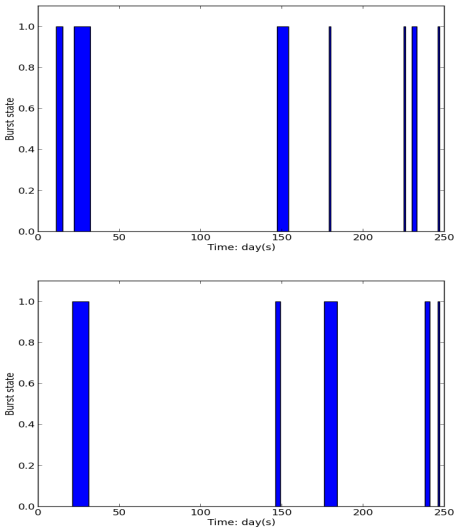

Figure 3: Burst states of "小威" (the upper) and "Serena Williams" (the lower) are correlated.

where $w(v)$ denotes the word of the node v and $\boldsymbol{s}_w$ denotes the burst sequence of the word w in which each entry is a binary variable indicating if w bursts at a moment throughout the time frame. Note that in the above equation, we regard $\boldsymbol{s}_{\boldsymbol{w}}$ as a vector. The numerator is the number of days when $w(c)$ and $w(e)$ co-burst and the denominator is the number of days when either $w(c)$ or $w(e)$ bursts.

3.4 Graph-based Decipherment

We define the overall (credibility) score as the linear combination of the clues introduced above:

$$Score(c, e) = \eta S_p + \lambda S_t + \gamma S_n + \delta S_b \quad (3)$$

where S_p, S_t, S_n and S_b are the scores that measure the value/reliability of the pronunciation, translation, neighbor and burst correlation clues respectively, and η, λ, γ and δ are hyperparameters for adjusting their weights.

Based on Eq (3), we can now compute the score of any candidate pair $\langle c, e \rangle$. For pairs that are known to be correct alignments according to prior knowledge, their overall scores will be fixed to 1.0. For other possible candidate pairs, we simply initialize their scores as follows:

$$Score(c, e) = \frac{1.0}{|Cand(c)|} \quad (4)$$

where $Cand(c)$ is the set of c's candidate nodes in the English BINet.

Given that $Score(c, e)$ is influenced by other pairs' scores, we design an iterative algorithm to compute and update the scores to decipher the entire Chinese BINet through propagation. This process is elaborated in Algorithm 1.

Algorithm 1 Graph-based Decipherment

1: For the determined pair $\langle c, e\rangle$ based on the prior knowledge, $Score(c, e) \leftarrow 1.0$
2: For other undermined pairs $\langle c, e\rangle$, initialize $Score(c, e)$ according to Eq (4);
3: **while** $True$ (until $\Delta Conf(G_c, G_e) \leq 0.0001$) **do**
4: **for** each undetermined pair $\langle c, e\rangle$ **do**
5: Compute new_score according to Eq (3);
6: $update(c, e) = \min(1.0, new\_score)$
7: **end for**
8: **for** each undetermined pair $\langle c, e\rangle$ **do**
9: $Score(c, e) \leftarrow update(c, e)$
10: **end for**
11: **end while**

$\Delta Conf(G_c, G_e)$ in the 3rd line of Algorithm 1 is the difference between the network decipherment confidence score at the current iteration and that at its previous iteration. $Conf(G_c, G_e)$ is defined as follows, reflecting how much confidence we have in our network decipherment result:

$$Conf(G_c, G_e) = \sum_{c \in V_c} \max_{e \in Cand(c)} Score(c, e) \quad (5)$$

In practice, propagation of prior knowledge and clues makes the confidence score increase because it helps us know more about the network (as illustrated by Figure 2). When the confidence score stops increasing or increases marginally (≤ 0.0001) after several iterations, the algorithm terminates[4].

4 Experiments

We first evaluate our approach on aligning nodes in the cross-lingual BINets for fine-grained cross-lingual stream alignment in Section 4.1. Then, we show the value of derived alignments for endless language knowledge acquisition in Section 4.2.

4.1 Stream alignment

4.1.1 Data

We used the public 2010 Agence France Presse (AFP) news in Chinese (Graff and Chen, 2005) and English Gigaword (Graff et al., 2003) as our cross-lingual text streams. The Chinese stream has 17,327 while the English one contains 186,737 documents.

We removed stopwords, conducted lemmatization and name tagging for the English stream, and did word segmentation and name tagging for the Chinese stream using the Stanford CoreNLP toolkit (Manning et al., 2014).

[4]Due to the upper bound of $Conf(G_c, G_e)$, the algorithm must terminate after several iterations.

We detected bursts and constructed the BINets[5] for the Chinese and English stream based on (Ge et al., 2016a). The constructed Chinese BINet has 7,360 nodes and 33,892 edges while the English one has 8,852 nodes and 85,125 edges. Our seed bi-lingual lexicon is released by (Zens and Ney, 2004), containing 81,990 Chinese word entries, each of which has an English translation. Among the 7,360 nodes in the Chinese BINet, 2,281 nodes need to be deciphered since their words are not in the bi-lingual lexicon.

4.1.2 Evaluation Setting

We evaluate our approach in an end-to-end fashion. For a node c in the Chinese BINet, we choose the node e^* which has the highest score as c's counterpart in the English BINet:

$$e^* = \arg \max_{e \in Cand(c)} Score(c, e)$$

We rank the aligned node pairs by the score and manually evaluate the quality of the top K pairs. A pair $\langle c,e\rangle$ is annotated as correct if e is a correct translation of c or e refers to an entity that c refers to. The annotation assignment is done by three human judges with 89.4% agreement. The disagreement mainly arises from the ambiguity of some named entities. In the evaluation, we consider $\langle c,e\rangle$ correct if more than two judges annotate it as correct.

We compare our approach to the following baselines that use various combinations of clues to verify candidates for decipherment as well as the state-of-the-art algorithm for language decipherment from non-parallel corpora:

- Pronunciation verification (*pv*): Use the pronunciation clue only
- Translation verification (*tv*): Use the translation clue only
- Neighbor verification (*nv*): Use the neighbor clue only to decipher the BINet through propagation.
- Correlation of burst verification: (*cv*): Use the burst correlation clue only
- *pv*+*tv* and *pv*+*tv*+*nv*
- *Bayesian Inference*: Bayesian inference based decipherment approach (Dou and Knight, 2012) based on the alignment of bigram language

[5]We discarded the edges whose weight is smaller than a threshold (5 for Chinese and 20 for English BINet given the difference of their data size) for removing trivial connections.

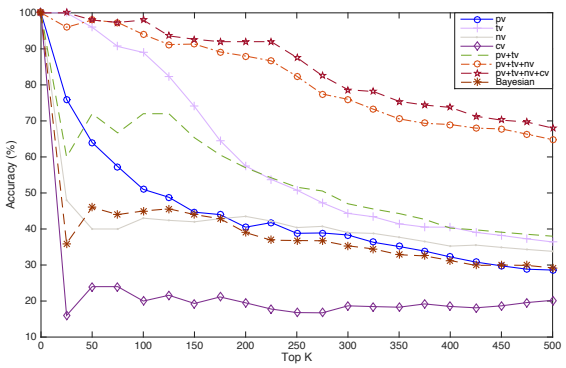

Figure 4: Accuracy curves of various approaches. Note that *pv+tv+nv+cv* is our final approach.

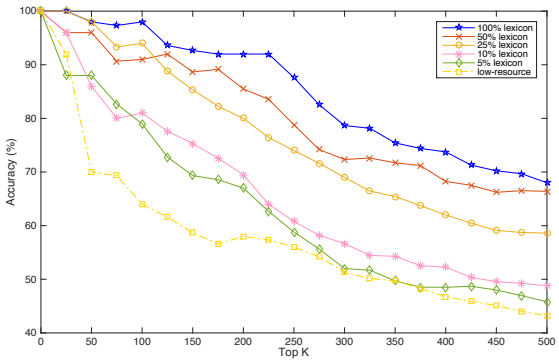

Figure 5: Accuracy curves of our approach with different resource settings.

models across languages. We adapt it to our experiment setting by considering adjacent nodes in a BINet as bigrams for decipherment.

We used 2009 AFP Chinese/English news in Gigaword as our development set to tune hyperparameters. Since our approach has only 4 parameters (i.e., η, λ, γ, δ in Eq (3)), it is easy to tune the parameters using grid search (from 0.0 to 1.0 with a step 0.2) on the development set. For baselines except Bayesian inference, the score computation function is almost identical to Eq (3) except that the weights of the clues which are not used are set to 0.

4.1.3 Results

We present the results in Figure 4. Our approach outperforms all the baselines because it considers various clues for decipherment. Among the baselines, accuracy scores of *pv* and *tv* drop dramatically with K increasing because a single clue can only decipher a limited number of nodes effectively. *pv+tv* seems to alleviate the problem to some extent: its accuracy does not drop so drastically as *pv* or *tv* because multiple clues allow us to decipher more nodes but its accuracy is still not desirable. Among the clues, *cv* performs worst, demonstrating that the burst correlation clue alone is far from enough for decipherment. Compared with *pv*, *tv* and *cv*, *nv* deciphers the nodes in the Chinese BINet through propagation but the neighbor clue alone is not sufficient for accurate decipherment. It is notable that *nv* achieves comparable performance to the Bayesian inference method which uses similar clues, demonstrating the effectiveness of our decipherment framework despite its simplicity. Moreover, our graph-based decipherment approach is more flexible to incorporate a variety of clues. When it is combined with *pv+tv*, the performance shows a significant boost and achieves approximately 90% accuracy in the top 200 results though it is slightly inferior to our final approach due to the lack of awareness of burst correlation.

Another interesting observation from Figure 4 is that our approach clearly know the confidence of its predictions. For top 100 mined pairs with the highest confidence scores (i.e., the score in Eq (3)), **the accuracy is 98%**. Therefore, it is easy to control the quality of mined pairs, which is important for a text mining algorithm.

We also study the effect of language resources on the performance. We first randomly sample different sizes of entries from the original bi-lingual lexicon as new bi-lingual lexicons. The results[6] in Figure 5 show the accuracy improves as the size of bi-lingual lexicon grows because more prior knowledge benefits deciphering the BINet. In addition, we test our approach in a low-resource setting where there is no knowledge of the romanization system (i.e., pinyin) and no pre-trained word segmentation and name tagging tools are available. The only available resource is a very small bi-lingual lexicon with 1,000 most common Chinese words[7] and their corresponding English translations. In this setting, we use an unsupervised Chinese word segmentation approach combining a Hierarchical Dirichlet Process (HDP) model with a Bayesian HMM model (Chen et al., 2014) to segment Chinese text instead of the preprocessing steps mentioned in Section 4.1.1. According to Figure 5, our approach still performs well in the low-resource setting although its accuracy curve is lower than that in rich-resource settings, demonstrating it can work in both rich- and low-resource settings.

[6]The sample processes are repeated for 3 times and the results are the averaged accuracy.

[7]We sample these Chinese words based on IDF.

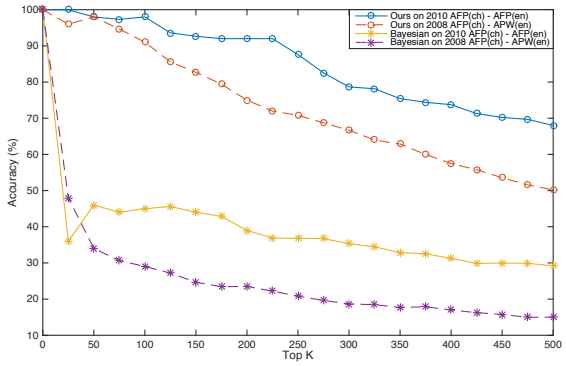

Figure 6: Accuracy curves on multiple datasets.

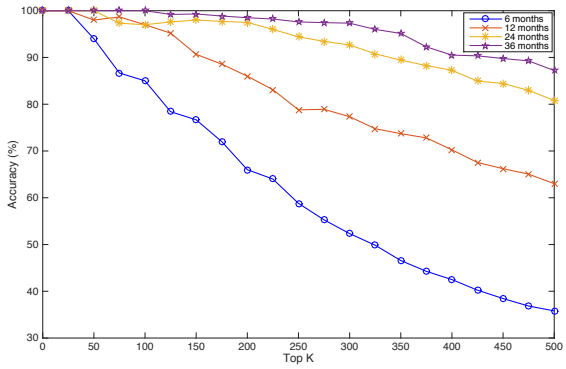

Figure 7: Stream alignments derived from 2008 on the Chinese and English AFP news streams. With more data (i.e., streams within a longer time frame) being aligned, our approach can harvest more high-quality alignments.

In order to test the generalization ability, we evaluate our approach using the same hyperparameters on another coordinated text streams – AFP Chinese and APW English news stream in 2008. The results in Figure 6 show that our decipherment approach consistently outperforms the other baseline and still deciphers the top 100 nodes in high accuracy even though the curve in 2008 is lower than that in 2010. The performance difference in 2008 and 2010 mainly arises from the difference on topic overlaps. In the streams of 2010, the Chinese and English news are from the same news agency (i.e., AFP). Therefore, the topic overlaps of the cross-lingual streams are larger than 2008, allowing more nodes to be deciphered correctly.

Finally, we investigated the performance of our approach under various sizes of data provided, as shown in Figure 7. As observed, when the data size is small (e.g., 6-month coordinated text streams), the approach works poorly because there are very few nodes in BINets that can be aligned. As the data size increases, our approach can efficiently[8] harvest a growing number of high-quality

[8]Efficiency is reported in the supplementary notes.

Node	Chinese (burst period)	English (burst period)
1	贝鲁斯柯尼 (346-348)	Berlusconi (344-348)
2	卫报 (332-337)	Guardian (334-336)
3	曼城 (360-363)	Manchester City (358-361)
4	鸠山 (147-158)	Hatoyama (147-158)
5	空域 (106-111)	airspace (104-111)
6	纾困 (119-130)	bailout (112-129)
7	国际奥林匹克委员会 (38-45)	IOC (36-44)
8	恋童癖 (102-104)	paedophilia (90-108)
9	小威 (147-153)	Serena Williams (146-148)
10	东协 (200-203)	ASEAN (198-203)
11	东南亚国家协会 (299-302)	ASEAN (299-303)
12	央行 (129-130)	European Central Bank (125-129)
13	央行 (75-76)	Bank of Japan (73-76)
14	翁山 (310-311)	Aung San Suu Kyi (308-313)
15	苏姬 (310-311)	Aung San Suu Kyi (308-313)

Table 1: Alignment examples. The numbers of burst periods denote the number of days after Jan 1, 2010.

alignments, as reflected by the higher curves in Figure 7. Considering massive coordinated text streams generated every day, if the approach can be applied to the endless streams, it is possible to monitor the streaming data and derive countless alignments for never-ending language knowledge acquisition.

4.2 Endless language knowledge mining

Table 1 shows the stream alignment result of our approach. As demonstrated above, we can derive a variety of language knowledge from the fine-grained cross-lingual alignments.

Word/entity translations are the main knowledge that can be derived from our alignment results by extracting word pairs from the aligned cross-lingual node pairs. Formally, we find a Chinese word w's English translation w^* as follows:

$$w^* = w(e^*)$$

$$e^* = arg\max_{e \in V_e} \max_{c \in V_c(w)} Score(c, e)$$

where $V_c(w)$ is the set of Chinese nodes whose word is w, and $w(e)$ denotes the word of node e.

We evaluate our approach on mining translations of bursty Chinese words, based on the evaluation criteria of bilingual lexicon extraction. Specifically, we test how many out-of-vocabulary (OOV) words appearing in the Chinese BINet are correctly translated. The datasets used for evaluation are the 2010 and 2008 streams in Figure 6. In total, there are 1,226 and 1,082 distinct Chinese OOV words (excluding incorrectly segmented words) in the corresponding Chinese BINets. Accuracy is used to measure the proportion of the words being correctly translated, as (Tamura et al., 2012) did.

Table 2 compares our approach to representative bilingual lexicon extraction approaches. CONTEXT is one of the earliest approaches for extracting word translations from comparable cor-

Model	Acc_1(2010)	Acc_1(2008)
CONTEXT (Fung and Yee, 1998)	0.32%	0.37%
COLP (Tamura et al., 2012)	0.32%	0.46%
SIMLP (Tamura et al., 2012)	0.49%	0.46%
DIVERSE (Schafer and Yarowsky, 2002)	5.22%	4.25%
DIVERSESP (Sproat et al., 2006)	5.46%	4.44 %
BAYESIAN(LM) (Dou and Knight, 2012)	0.57%	0.55%
BAYESIAN(BINET)	11.17%	4.81%
Ours	**28.38%**	**19.78%**

Table 2: Performance of translating bursty words.

pora based on context similarity. COLP and SIMLP are label propagation models on word co-occurrence and similarity graphs for bilingual lexicon extraction. DIVERSE is a variant of CONTEXT by adding various information (e.g., pronunciation and temporality) and DIVERSESP is the approach using phonetic and frequency correlation with a score propagation strategy. BAYESIAN is the Bayesian decipherment approach which has been introduced in the previous section, and it is evaluated in two settings (i.e., based on traditional bigram language models and BINets). According to Table 2, our approach substantially outperforms the other approaches on both datasets, showing its advantages for mining translation of bursty words in coordinated text streams. It is also notable that the BINet-based BAYESIAN improves the LM-based counterpart, demonstrating the advantage of burst-level alignment for this task.

In addition to the comparisons to the classical baselines, we also test the latest representative unsupervised bi-lingual lexicon extraction approaches (Zhang et al., 2017a,b) based on word embedding and generative adversarial nets (GANs). Unfortunately, these approaches do not perform well in our setting. For example, the approach in (Zhang et al., 2017a) achieved <1% accuracy[9]. One reason is that the topic overlap of coordinated cross-lingual text streams is not so significant as the Wikipedia data used for their experiments, and the other reason is that their approaches focus on common fundamental words like "城市(city)" while our targets are OOVs like "东协(ASEAN)" which do not frequently appear in a corpus. In contrast, our approach is more practical: it not only works well in easily available and endless coordinated text streams without high content overlap requirement, but also can accurately mine translations of many OOVs which do not appear frequently and really need mining their translations.

[9]We implement this approach using the codes released by the authors. Their reported accuracy for the common words with over 1,000 occurrences is 2.53% on Gigaword corpus.

Knowledge	Derived	Correct	Acc
Word/entity translation pair	500	416	0.83
Polysemy/multi-referential word	11	8	0.73
Synonym/Co-referential word pair	72	49	0.68
Entity phrase	99	84	0.85

Table 3: Knowledge derived from top 500 alignments obtained by aligning AFP Chinese and English text streams from 2002 to 2010.

As illustrated in Figure 1, besides word/entity translations, various types of knowledge can also be derived from the BINet alignment results as by-products. For example, for node 9 in Table 1, deciphering the nickname "小威" into *Serena Williams* can benefit cross-lingual entity linking. Nodes 10-11 also demonstrate the potential effect on synonym detection, entity linking and coreference resolution, like the case of Figure 1(b). Nodes 12-13 show that the deciphered BINets can detect polysemous/multi-referential word like "央行(Central bank)" which may refer to different entities during different burst periods, like Figure 1(c). Moreover, the deciphered BINets can also help entity phrase extraction based on the idea of Figure 1(d). For example, in nodes 14-15, 翁山苏姬(Aung San Suu Kyi) is not recognized as a person name by the Chinese name tagger; instead, it is mistakenly separated into two words – 翁山(Aung San) and 苏姬(Suu Kyi). However, since 翁山(Aung San) and 苏姬(Suu Kyi) are deciphered into the same English named entity – Aung San Suu Kyi, we can merge them back to form the correct entity.

For evaluating our approach's performance on language knowledge acquisition, we align the AFP Chinese-English text streams from 2002 to 2010. The Chinese stream has 119,196 documents and the English one contains 1,608,636 documents. Our approach obtained 7,211 node alignments[10]. Among them, we focus on the top 500 alignments to guarantee their quality and use the aforementioned idea for deriving language knowledge.

Table 3 shows the result of deriving knowledge from the alignments. Among top 500 alignments, we derived 416 correct word/entity translation pairs with 83% accuracy. Also, we correctly derived 8 polysemous/multi-referential words, 49 synonymous/co-referential word pairs and 84 entity phrases as byproducts. It is notable that the data size of coordinated cross-lingual text streams available on the web is much larger than that used in our experiment and they are endlessly updated.

[10]The alignments with a low score (< 0.05) are discarded.

That means it is promising to endlessly derive language knowledge by applying our approach to the huge size of endless cross-lingual text streams, which may benefit NLP applications like machine translation, entity linking and name tagging.

5 Related Work

Previous studies on cross-lingual text stream alignment tend to focus on coarse-grained (i.e., topic-level) alignment for finding common patterns (Wang et al., 2007; De Smet and Moens, 2009; Wang et al., 2009; Zhang et al., 2010; Hu et al., 2012) and discovering parallel sentences and documents (Munteanu and Marcu, 2005; Enright and Kondrak, 2007; Uszkoreit et al., 2010; Smith et al., 2010; Krstovski and Smith, 2011, 2016) across languages. Studies on fine-grained cross-lingual alignment are mainly for bilingual lexicon induction (e.g., (Fung and Yee, 1998; Rapp, 1999; Koehn and Knight, 2002; Schafer and Yarowsky, 2002; Shao and Ng, 2004; Schafer III, 2006; Hassan et al., 2007; Haghighi et al., 2008; Udupa et al., 2009; Klementiev and Callison-Burch, 2010; Tamura et al., 2012; Irvine and Callison-Burch, 2013, 2015b; Kiela et al., 2015; Irvine and Callison-Burch, 2015a; Vulic and Moens, 2015; Cao et al., 2016; Zhang et al., 2017b,a)) and name translation mining (e.g., (Sproat et al., 2006; Klementiev and Roth, 2006; Udupa et al., 2008; Ji, 2009; won You et al., 2010; Kotov et al., 2011; Lin et al., 2011; Sellami et al., 2014)) from non-parallel corpora. However, these approaches are mainly developed for general comparable corpora, not specially for cross-lingual text streams; thus many of them did not use the powerful stream-level information (e.g., co-burst across languages). In contrast to the word-level alignment methods, we attempt to mine burst-level alignment to largely narrow down candidates, and introduce powerful clues for improving accuracy and discovering various language knowledge.

In contrast to previous cross-lingual projection work like data transfer (Pado and Lapata, 2009) and model transfer (McDonald et al., 2011), we do not require any parallel data. Moreover, our BINets are cheap to construct, which can be easily extended to other languages. This is also the first attempt to apply the decipherment idea (e.g., (Ravi and Knight, 2011; Dou and Knight, 2012; Dou et al., 2014)) to graph structures instead of sequence data.

6 Conclusions and Future Work

This paper proposes an approach to deciphering the Burst Information Network constructed from foreign languages as a novel way to align cross-lingual text streams. For the first time we propose to model stream alignment as a network decipherment problem. By leveraging the network structures with stream-level burst features as well as various clues, our approach can accurately align the important information units across languages and derive a variety of knowledge. Given that our approach is unsupervised, effective, intuitive, interpretable, and easily implementable, it is promising to use it as a framework for never-ending language knowledge mining from big data, which might benefit NLP applications such as machine translation and cross-lingual information access.

For future work, we plan to 1) conduct more experiments and analyses following this preliminary study to verify our approach's effectiveness for more languages and domains (e.g., social stream VS news stream); 2) attempt to use word embedding (e.g., word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018)) for local context encoding and use it as a clue for decipherment; 3) apply our approach to real-time coordinated text streams for never-ending knowledge mining and use the mined knowledge to improve the downstream applications.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. We also want to thank Xiaoman Pan, Dr. Taylor Cassidy, Dr. Clare R. Voss, Prof. Jiawei Han, Prof. Sujian Li and Prof. Yu Hong for their helpful comments and discussions. This work is supported by NSFC project 61772040 and 61751201. Heng Ji's work has been supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014, Air Force No. FA8650-17-C-7715, ARL NS-CTA No. W911NF-09-2-0053 and NSF Awards IIS-0953149 and IIS-1523198. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. The contact author is Zhifang Sui.

References

Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. 2016. A distribution-based model to learn bilingual word embeddings. In *COLING*.

Miaohong Chen, Baobao Chang, and Wenzhe Pei. 2014. A joint model for unsupervised chinese word segmentation. In *EMNLP*.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *ACM workshop on Social web search and mining*.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *EMNLP*.

Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *EMNLP*.

Jessica Enright and Grzegorz Kondrak. 2007. A fast method for parallel document identification. In *NAACL*.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel and comparable texts. In *COLING-ACL*.

Tao Ge, Lei Cui, Baobao Chang, Sujian Li, Ming Zhou, and Zhifang Sui. 2016a. News stream summarization using burst information networks. In *EMNLP*.

Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, and Ming Zhou. 2016b. Event detection with burst information networks. In *COLING*.

Tao Ge, Lei Cui, Heng Ji, Baobao Chang, and Zhifang Sui. 2016c. Discovering concept-level event associations from a text stream. In *NLPCC*.

David Graff and Ke Chen. 2005. Chinese gigaword. *LDC Catalog No.: LDC2003T09, ISBN*, 1:58563–58230.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In *RANLP*.

Shuo Hu, Takahashi Yusuke, Liyi Zheng, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, and Yoji Kiyota. 2012. Cross-lingual topic alignment in time series japanese/chinese news. In *PACLIC*.

A. Irvine and C. Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *NAACL*.

Ann Irvine and Chris Callison-Burch. 2015a. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 1(1).

Ann Irvine and Chris Callison-Burch. 2015b. Discriminative bilingual lexicon induction. *Computational Linguistics*, 1(1).

Heng Ji. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*.

Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *EMNLP*.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Alexandre Klementiev and Chris Callison-Burch. 2010. Bilingual lexicon induction for low-resource languages. In *JHU Technical Report*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *COLING-ACL*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL workshop on Unsupervised lexical acquisition*.

Alexander Kotov, ChengXiang Zhai, and Richard Sproat. 2011. Mining named entities with temporally correlated bursts from multilingual web news streams. In *WSDM*.

Kriste Krstovski and David A Smith. 2011. A minimally supervised approach for detecting and ranking document translation pairs. In *WMT*.

Kriste Krstovski and David A Smith. 2016. Bootstrapping translation detection and sentence extraction from comparable corpora. In *NAACL*.

Qi Li, Heng Ji, Yu Hong, and Sujian Li. 2014. Constructing information networks using one single model. In *EMNLP*.

Wen-Pin Lin, Matthew Snover, and Heng Ji. 2011. Unsupervised language-independent name translation mining from wikipedia infoboxes. In *EMNLP Workshop on Unsupervised Learning for NLP*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language ssing toolkit. In *ACL: System Demonstrations*.

R. McDonald, S. Petrov, and K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

S. Pado and M. Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *ACL*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *ACL*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*.

Charles F Schafer III. 2006. *Translation discovery using diverse similarity measures*. Johns Hopkins University.

Rahma Sellami, Fatiha Sadat, and Lamia Belguith Hadrich. 2014. Mining named entity translation from non parallel corpora. In *FLAIRS*.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *COLING*.

Jason R Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *NAACL*.

Richard Sproat, Tao Tao, and ChengXiang Zhai. 2006. Named entity transliteration with comparable corpora. In *ACL*.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *EMNLP*.

Raghavendra Udupa, K Saravanan, A Kumaran, and Jagadeesh Jagarlamudi. 2008. Mining named entity transliteration equivalents from comparable corpora. In *CIKM*.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In *EACL*.

Jakob Uszkoreit, Jay M Ponte, Ashok C Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *COLING*.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL*.

X. Wang, C. Zhai, X. Hu, and R. Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *KDD*.

Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. 2009. Mining common topics from multiple asynchronous text streams. In *WSDM*.

Gae won You, Seung won Hwang, Young-In Song, Long Jiang, and Zaiqing Nie. 2010. Mining name translations from entity graph mapping. In *EMNLP*.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *HLT-NAACL*.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *ACL*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *EMNLP*.