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

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

1 Burst Information Network Decipherment

We elaborate the process of Burst Information Network decipherment in this section to help better understand the idea of our approach.

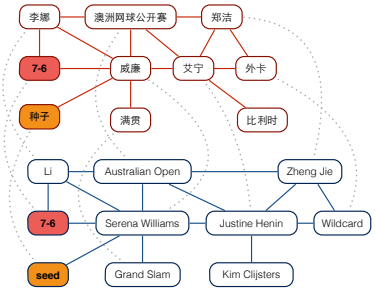

Figure 1: Alignments based on the prior knowledge.

Figure 1 shows the nodes that are deciphered based on the prior knowledge. For example, "7-6" is a language-universal representation and thus can be easily deciphered. "种子(seed)" is a basic Chinese word which should be included in almost all the general Chinese dictionaries. By using a bilingual lexicon as prior knowledge, it can be properly deciphered as "seed" in English without much effort.

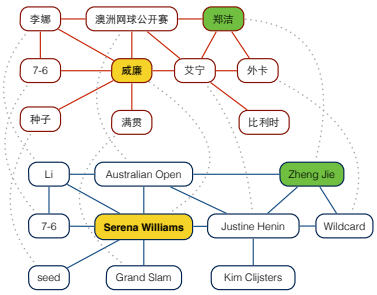

Figure 2: Alignments based on the pronunciation clue.

For some nodes that cannot be deciphered by the prior knowledge, we can use various clues to decipher them. For example, Figure 2 and 3 illustrates some example nodes that are deciphered by the pronunciation clue and the translation clue respectively.

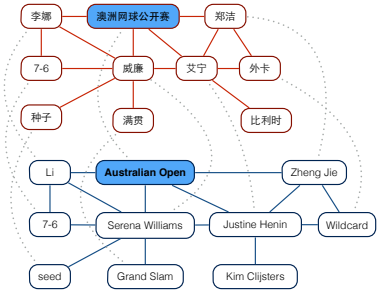

Figure 3: Alignments based on the translation clue.

In addition, some nodes can be deciphered by the alignment of their adjacent nodes. For example, in Figure 4(a), we have some nodes that have already been deciphered with the prior knowledge and some clues. These deciphered nodes can help decipher their adjacent nodes such as "艾宁(Henin)", as shown in Figure 4(b). When "艾宁" is deciphered as "Henin", such knowledge can further help decipher its adjacent node "外卡(wildcard)" (Figure 4(c)).

Throughout this paper, the time unit (of burst periods) is one day.

2 Experiments

In this section, we discuss the comparison to baselines, analyze the experimental results by comparing to name transliteration, report the performance on model efficiency and give the implementation details of deriving language knowledge.

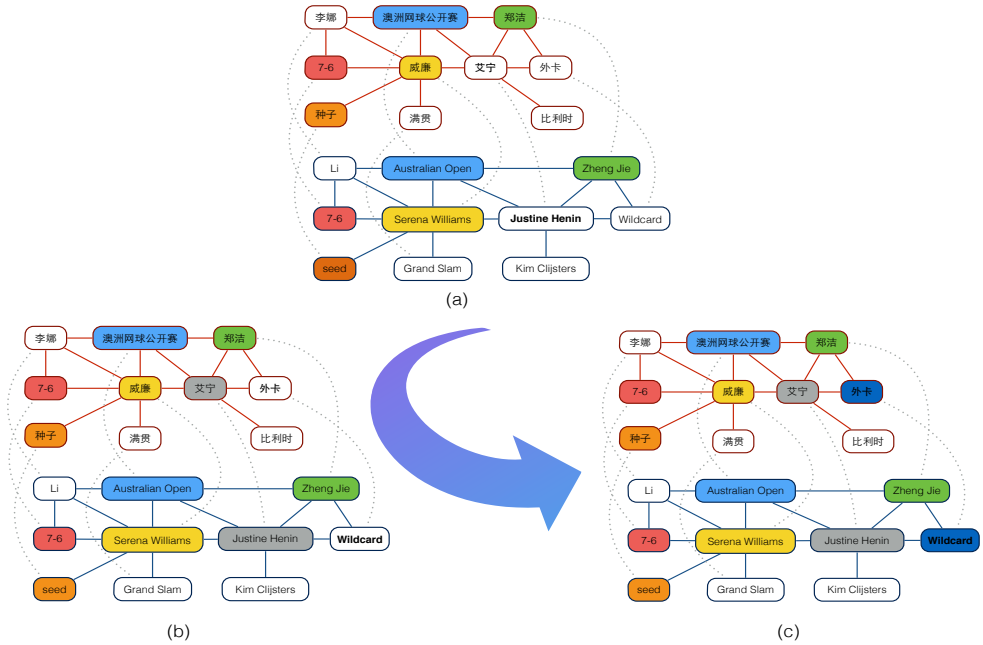

Figure 4: Deciphering the network based on the neighbor clue in a propagation manner.

2.1 Baselines in our experiments

In the paper submission, we report some combinations of the clues. It is notable that there are totally 15 combinations of the clues and we cannot list all their performance in the paper submission – We only report a few of them in the paper submission because the performance of all the others is inferior to *pv*+*tv*+*nv*+*cv*.

In the paper submission, we report the comparison to some representative approaches on bilingual lexicon extraction. It is notable that although some other previous studies (e.g., (Kotov et al., 2011)) are related to ours, we did not compare to them because their focuses and settings are different from ours. For example, the focus of (Kotov et al., 2011) is to better compute/detect burst correlation that can be used for name transliteration mining, while ours is the design of the original paradigm and framework of accurate fine-grained stream alignment for mining various language knowledge endlessly. In our work, burst correlation only works as a clue for decipherment, which is simply computed by equation (2) in the paper submission because our focus is not computing it. Second, the setting of (Kotov et al., 2011) is different from ours. Since their focus is burst correlation computation instead of the task of mining name transliteration (their paper's keywords even DON'T include name transliteration), they mined name transliteration only based on burst correlation to compare to other work that studies burst/frequency correlation computation in the same setting. In fact, name transliteration mining is more than burst correlation computation, which needs various information and techniques for good performance. Moreover, the scope of the task we study is far beyond name transliteration (let alone burst correlation computation), which will be discussed in the following subsection.

The baselines are re-implemented for comparing to our approach on the coordinated text streams based on the descriptions of their original papers or the implementation of their released source codes or softwares.

2.2 Comparison to name transliteration

We analyzed translation pairs mined by our approach to see how many of them can be obtained by a transliteration model which is often used for name translation. Among top 100 translation pairs, only 9% can be correctly transliterated by a transliteration model (Jiampojamarn et al., 2007), demonstrating that our approach can discover large numbers of translation pairs that cannot be transliterated.

2.3 Efficiency

Table 1 shows the run time of our decipherment algorithm on different sizes of streaming data, which is measured on a workstation with Intel Xeon 3.5

	First 6 months			Last 6 months			1 year		
	#Node	#Edge	#Doc	#Node	#Edge	#Doc	#Node	#Edge	#Doc
Chinese	3,592	17,435	8,394	3,171	12,862	8,933	7,360	33,892	17,327
English	5,078	28,326	114,159	2,948	43,473	72,578	8,852	85,125	186,737
Time	161.27s			480.96s			979.72s		

Table 1: Run time of the decipherment algorithm on different sizes of streaming data (2010 Chinese-English AFP news articles).

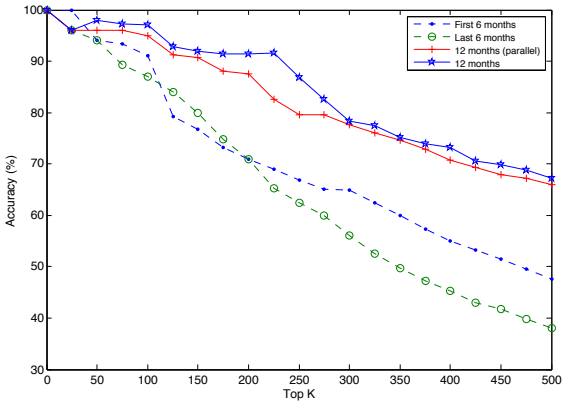

Figure 5: Performance of decipherment algorithm in parallel

GHz CPU and 64GB RAM. As data increases, it will take longer time for decipherment. However, it is notable that the decipherment processes for the data of the first 6 months and the last 6 months are independent and thus it is possible to run the decipherment algorithm on these two datasets in parallel and merge the decipherment results.

Figure 5 shows the performance of the decipherment algorithm in the parallel fashion. We can see that deciphering in parallel does not result in a significant decrease of accuracy. Therefore, we can split a text stream into several small parts and decipher them in parallel, which makes the decipherment more efficient. For the example in Table 1, if we decipher the text streams of the first 6 months and the last 6 months in parallel, the run time of decipherment would be 480.96 seconds assuming the time for merging the results is negligible.

2.4 Language knowledge derivation

The paper submission has introduced the basic idea for deriving language knowledge from the cross-lingual BINet node alignments. However, there is some details that should be taken into account for a better result, especially for polysemous word knowledge discovery.

Basically, if one Chinese word is aligned to multiple English words during different periods, then we consider it as a polysemous word. However, there is some cases that the multiple English words are synonyms. For example, we detect the Chinese word "中情局(Central Intelligence Agency)" is aligned to two English words (*Central_Intelligence_Agency*; *CIA*) at different time but we cannot say "中情局" is a polysemous word because *CIA* is just abbreviation of *Central Intelligence Agency*. For avoiding such a case, we use word embedding to tell if the multiple English words have high semantic similarity (> 0.9). Specifically, we use the whole English Gigaword corpus to train the word embedding using the toolkit word2vec[1] (Mikolov et al., 2013).

2.5 More languages

We also conducted preliminary experiments on Japanese-English news stream from Bing news[2]. The time frame of the coordinated streams is from February 5 to December 31 in 2015. The number of news articles in Japanese and English streams are 8.2M and 75.2M respectively. We used the unsupervised word segmentation approach (mentioned in Section 4.1.3 in our paper) to segment Japanese texts and extract Romaji for Hiragana and Katakana as pronunciation features. We used JMdict[3] as the Japanese-English bi-lingual lexicon, which has approximately 40K entries.

Among top 100 mined pairs, there are 44 pairs that are annotated correct. The accuracy is inferior to that of Chinese-English Gigaword news stream. The main reason for that is that the topic overlap (especially burst topic overlap) in Japanese-English Bing news streams is very little. For Bing news stream, Japanese news tends to report local events in Japan while English news tends to report international or local events in English speaking countries like US. Therefore, for Japanese burst words like ニコニコ[4], it is almost impossible to find its counterpart in the English stream. As a result, among the mined pairs, there are only 7 pairs with high scores (≥ 0.80) yet all of them are cor-

[1] https://code.google.com/archive/p/word2vec/
[2] https://www.bing.com/news
[3] http://www.edrdg.org/jmdict/j_jmdict.html
[4] A Japanese video website: http://www.nicovideo.jp/

rect; while in the Chinese-English news stream, there are more than 30 mined pairs whose scores are higher than 0.8 (also in almost 100% accuracy). It is also noted that Japanese words derived by our unsupervised word segmentation approach often fail to match the entries in the lexicon. For example, we got a word "に関連" from the word segmentation model. It cannot match the entry "関連" in the lexicon, which impairs the functional value of the bi-lingual lexicon. Therefore, we think the performance should be further improved if a supervised or dictionary-based word segmentation model can be applied. In addition, a Japanese word usually can be written in multiple forms (e.g., Kanji, Hiragana, Katakana or their mix), which is also a challenge for stream alignments if no text normalization is performed. We will investigate more on stream alignment in terms of these kinds of challenges for more languages as our future work.

References

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *NAACL*.

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

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