Evaluation and Comparison of Cross-lingual Text Processing Pipelines

Robert Jungnickel, André Pomp, Andreas Kirmse, Xiang Li and Vladimir Samsonov
Institute of Information Management in Mechanical Engineering, RWTH Aachen University, Aachen, Germany
Email: {robert.jungnickel, andre.pomp, andreas.kirmse, xiang.li, vladimir.samsonov}@ima.rwth-aachen.de

Tobias Meisen
Chair of Technologies and Management of Digital Transformation, University of Wuppertal, Wuppertal, Germany
Email: meisen@uni-wuppertal.de

***Abstract*—With the trend of globalization and digitalization, many transnational companies are continuously collecting and storing unstructured text data in different languages. To exploit the business value of such high-volume multilingual text data, cross-lingual information extraction utilizes machine translation and other natural language processing (NLP) techniques to analyze this data. However, results of these analysis heavily depend on the order in which the tasks are performed as well as the used machine translation and NLP approaches or trained models. In this paper, we defined and evaluated a series of cross-lingual text processing pipelines for English and Chinese language. We therefore combine multiple commercial machine translation services with different automatic keyphrase extraction and named entity recognition techniques and evaluate their performance with regards to the order of execution. Hence, we evaluate the combination of machine translation systems and natural language processing techniques with two processing sequences in our experiment. One is to translate the document before extracting keyphrase and named entities. The other is to translate the processing results. The experiment outcomes indicate that translating documents is a better choice than the other way around in both tasks. However, there exists a substantial disparity between the performance of the cross-lingual text processing pipelines and the corresponding monolingual references.**

***Keywords*—natural language processing; keyphrase extraction; translation; cross-lingual text processing; processing pipelines.**

I. Introduction

Natural language processing techniques such as keyphrase extraction and named entity recognition are utilized to facilitate automatic information extraction and analyses of text data. Such techniques are mostly applied in monolingual applications. However, in this highly globalized world, we often have text data that is written in different languages originating from different countries. The naive and often done solution is to define one base language, usually English, and translate everything into this base language before doing analysis. In this case, we are facing the problem of choosing between integrating analysis results and jointly analyzing text data in different languages. To resolve this problem, combining machine translation with other natural language processing techniques is a possible solution. In this paper, we refer to the systems which combine machine translation and other natural language processing techniques as cross-lingual text processing pipelines [1]. Combining machine translation and other natural language processing techniques differently can have different effectiveness because the translation quality varies from short phrases to long text and the natural language processing techniques usually perform differently on one language than another. We choose the language pair of Chinese and English to experiment, as the Chinese language has the most native speakers in the world and English is commonly treated as an universal language wordwide. The main question of this paper is how to join machine translation systems and other natural language processing techniques to retrieve the best result.

Similar to the cross-lingual text processing area, the Cross-Language Information Retrieval (CLIR) has the same goal of combining Information Retrieval and translation processes. CLIR is an established domain since the 1960s which enables users to query with one language and retrieve documents or information in another language. For CLIR systems the question is: should the documents be translated for querying or should the queries be translated to match the documents? Analog to CLIR, the cross-lingual text-processing pipelines also have two methods to couple the machine translation systems and other natural language processing techniques. One approach is to translate all text data before analyses; another approach is to translate the results after analyses. The choice between two possible approaches can be summarized into a sequencing problem of machine translation module and text processing module. However, only a few works discuss this processing sequence problem for cross-lingual text processing pipeline. For instance, Aone et al. [2] come up with the concept of a hybrid system of information extraction and machine translation for Japanese-English, which is similar to a cross-lingual text processing pipeline. This work lacks quantitative analysis and comparison. To contribute to the aforementioned research in a quantitative perspective , in this paper, we evaluate and compare the performance of processing sequences with cross-lingual keyphrase extraction and named entity recognition tasks.

The contributions of our paper is as follows:

Author manuscript

Final version of manuscript ©2019 IEEE
Presented at the 2019 IEEE Symposium Series on Computational Intelligence (SSCI) December 6-9 2019, Xiamen, China

- We demonstrate that processing sequence have substantial influence on the performance of cross-lingual text processing pipelines, and translating the text before extracting keyphrases and named entities has better overall performance than translating the keyphrases and named entities afterwards.
- We compare the performances of different commercial machine translation systems in cross-lingual text processing scenarios.
- We show that there exists a significant performance disparity between monolingual and cross-lingual keyphrase extraction / named entity recognition.

II. Related Work

Many research domains in Natural Language Processing have their cross-lingual subdomains which are related to the topic of this paper. These related fields include Cross-Language Information Retrieval, Cross-lingual information extraction, and Cross-lingual Summarization.

A. Cross-Language Information Retrieval

Aforementioned Cross-Language Information Retrieval (CLIR) is a domain that shares similar goals and problems with the area covered in this paper. Query translation is one of the common methods to couple query and documents written in different languages. Hull and Grefenstette's research [3] describes a dictionary-based approach to translate query according to machine-readable transfer dictionaries, which are manually or automatically converted from a bilingual French-to-English dictionary. Their experiment result reveals that query translation with word-based dictionary performs poorly on querying phrasal expressions, whereas the manually edited phrase dictionary can significantly boost the CLIR effectiveness. To resolve the phrasal translation and translation ambiguity problems, Corpora-based query translation [4] use parallel or comparable corpora to translate the query phrase. However, the size of the parallel corpora is a limitation of this system, because the source of parallel corpora is often limited in particular languages and domains. Another major query translation system is machine-translation-based query translation. Wu et al. [5] present a result that machine-translation-based query translation system not only works for long queries translation, but also excels in translating short queries.

In more recent research [6] [7], internal representation of machine translation process is utilized to effectively enhance the query translation CLIR performance.

Document translation is the opposite approach of query translation to bridge the retrieval between the query and documents in two languages. This approach translate all documents in the collection into the target language with machine translation system to enable the information retrieval. Some research [1], [8] shows the performance of document translation is comparable or better than query translation, however, document translation is more computationally expensive than query translation, since the text size of large document base is significantly larger than the size of queries.

B. Cross-lingual Information Extraction

To solve the translation problem in Cross-lingual Information Extraction (CLIE), Aone et al. [2] describe a hybrid approach to couple information extraction and machine translation. They propose two possible configurations of integrating information extraction and machine translation. The first approach conducts the information extraction on the document and then uses machine translation to convert the extracted result. The other approach translates the document before the information extraction. However, only the first approach is tested and no quantitative analysis is given in this work. Furthermore, their approach focuses on Japanese-English language, which has different characteristics than Chinese.

Sudo et al. [9] quantitatively evaluated two CLIE systems. The results show that the result translation system has significantly better precision and recall than the document translation system. They conclude that the performance difference between the two systems depends not only on the machine translation quality but also on the text processing algorithms in different languages.

More recent research [10] [11] [12] shows that the result translation system and the document translation system perform about the same in the 5W CLIE task [13].

C. Cross-lingual Automatic Summarization

For Cross-lingual summarization, Torres-Moreno [14] compared the document translation and the result translation. He empirically favors the concept of translating the generated summarization, because the machine translation errors may affect the following automatic summarization process. For actual implementation of Cross-lingual Text Summarization, Lin [15] proposed a cross-lingual document retrieval and summarization system MuST [15] which adopts the approach of translating the generated summarization. Evans et al. [16] and Huang et al. [17] approach this problem the other way around by translating multilingual documents into English before automatic summarization.

To the best of our knowledge, no quantitative comparison between the document translation and summarization translation can be found in cross-lingual summarization research.

The related studies mentioned in this section provide no conclusive answer to the question, which processing sequence performs better in cross-lingual text processing pipelines. Translating documents and translating results both have supporting arguments and are implemented in different applications.

III. Cross-Lingual Text Processing

In this section, we introduce the methodology of building and evaluating the cross-lingual text processing pipelines. We first cover the machine translation systems used in this paper. Before we then introduce the two text processing sequences in Section 3.2. The cross-lingual keyphrase extraction and

Authors manuscript

named entity recognition are detailed in Section 3.4 and 3.5, respectively.

A. Machine Translation

In this paper, the machine translation systems have two responsibilities: one is dataset preprocessing for the experiment, the other is translation module for cross-lingual text processing. We compared and evaluated four machine translation systems in this paper, namely *Youdao*, *Google*, *Baidu*, and *Bing*, in cross-lingual text processing pipelines.

As one of our experiment goal is to evaluate the mainstream machine translation systems such as Google, Baidu, and Bing translator, we use the relatively minor *Youdao* translator for the translation preprocessing of the dataset, which will be discussed in detail in Section IV-A We include *Youdao* translator in the test for reference, as other machine translation engines could have the same effect on the evaluation results if using them to preprocess the dataset.

B. Processing Sequences

Identical to the related works, two possible sequences are available for connecting the machine translation systems and the text processing techniques as demonstrated in Figure 1. One solution is translating the original document into the target language and then extracting keyphrases and named entities from the translated document. The other possible solution is first extracting keyphrases and named entities from the original text and then translating the keyphrases and named entities into the target language. In this paper, for simplicity, we refer to the first sequence as "machine-translation-first" and the latter as "machine-translation-later" .

C. Keyphrase Extraction

Since the machine translation process has equal effects on all algorithms, it is reasonable to speculate that the performance of an algorithm in monolingual keyphrase extraction task can indicate how well this algorithm will perform in a cross-lingual keyphrase extraction task.

However, the performance of an algorithm in cross-lingual keyphrase extraction task is not necessarily the same as in the monolingual scenario. In fact, the performance of algorithms is likely to be affected by the machine translation process and the linguistic difference between different languages. Thus, one of the research questions in this paper is for cross-lingual text processing pipelines, which keyphrase extraction algorithm is preferable.

To answer this research question, we select three representative keyphrase extraction algorithms to test and compare in cross-lingual text processing pipelines. The first algorithm is *TF-IDF* [18], for it is a widely used baseline for keyphrase extraction task. The second algorithm is *Kea* [19] which is a representative supervised keyphrase extraction algorithm. The third algorithm we select is *MultipartiteRank* [20], as it is a state-of-the-art graph-based keyphrase extraction algorithm. In this paper, we use the open-source Python-based keyphrase extraction toolkit *pke* [21] to test the three algorithms.

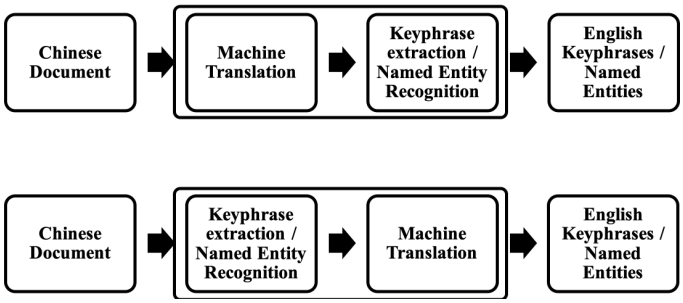

Fig. 1. Two possible processing sequences of cross-lingual text processing pipelines

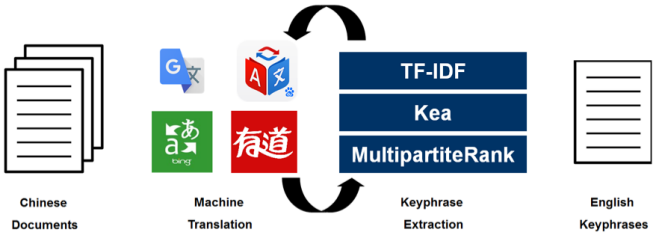

Fig. 2. Cross-lingual keyphrase extraction

To test the performance of cross-lingual keyphrase extraction pipelines, we have four options for machine translation systems, three keyphrase extraction algorithms, and two processing sequences, which provide us 24 pipelines to evaluate as Figure 2 illustrates. The goal of constructing all possible pipeline variations is to make sure that each influential factor is the single variate in a specific series of pipelines, which allows the independent evaluation of each influential factor in the cross-lingual keyphrase extraction pipelines.

As presented in Figure 2, due to two possible processing sequences, the keyphrase extraction algorithms need to process both English and Chinese documents in the experiment, which may potentially bring bias in the evaluation. For example, if the Kea model used on Chinese document is trained with large and high-quality training set while the English model is trained with a relatively limited resource, then it is quite possible that the Chinese Kea model performs better than the English one. In this case, it is almost certain that the pipelines which apply the keyphrase extraction to Chinese text (machine-translation-first) will perform better than the pipelines with English keyphrase extraction models (machine-translation-later). Here, the bias caused in model-training can affect the judgment of the influence of processing sequences.

To eliminate this type of bias in the experiment as much as possible, the English and the Chinese Kea models used in the experiment are trained respectively with the English version and the Chinese version of the *ScienceIE* dataset which is preprocessed as described in Section IV-A. The two versions of *ScienceIE* dataset are also used for *TF-IDF* to compute the Inverse Document Frequency values for English and Chinese keyphrase extraction for the same reason. The

Authors manuscript

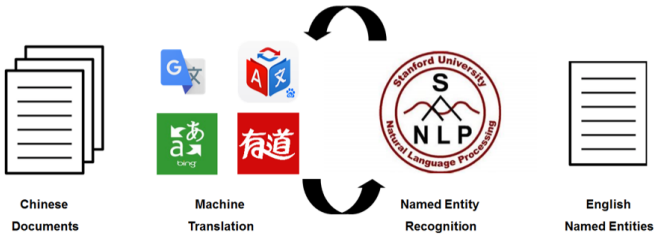

Fig. 3. Cross-lingual named entity recognition

MultipartiteRank algorithm requires no training process or reference document collection due to its unsupervised and graph-based character. However, the selection of topic and keyphrase candidates in the MultipartiteRank algorithm requires POS (Parts-of-Speech) information of the text. The English and the Chinese POS-taggers required for MultipartiteRank are from Stanford Natural Language Processing Group[22].

Another relevant research question for the cross-lingual keyphrase extraction task is how wide the gap between the performance of monolingual keyphrase extraction and cross-lingual keyphrase extraction is or if it is non existing. To assess the performance disparity, we test the three algorithms with the *ScienceIE* dataset without any translation as the monolingual performance reference.

D. Named Entity Recognition

In this paper, the evaluation of the cross-lingual named entity recognition pipelines focuses on the performance difference caused by various influential factors and the disparity between monolingual and cross-lingual named entity recognition. Three named entity categories, namely Person, Location, and Organization, are separately evaluated for a deeper understanding of the performance of the cross-lingual named entity recognition pipelines. With one named entity recognizer, two pipeline sequences, and four machine translation systems, we have in total eight pipelines to evaluate as Figure 3 illustrates. We use the named entity recognizer from Stanford coreNLP [22] toolkit in this paper for its compatibility for both English and Chinese.

IV. Experiment

In this section, we introduce the experiment setup and the evaluation results, with which we answer the related research questions.

A. Datasets and Preprocessing

Two datasets, namely ScienceIE dataset and Weibo Named Entity dataset are used in this paper. In this subsection, we introduce the three datasets and the corresponding preprocessing.

ScienceIE is a shared task in SemEval 2017 which is the conference for evaluations of computational semantic analysis systems in 2017. The goal of ScienceIE task is to extract information including keyphrases and relations from scientific publications. ScienceIE task provides a dataset which consists of 500 scientific documents evenly distributed among the domains Computer Science, Material Sciences, and Physics. In this dataset, 350 documents are training set, 50 documents are the development set, and the rest are test set. Following is a fragment in the ScienceIE dataset and the corresponding labeled keyphrases.

Document Fragment: *"Contact methods have been developed and used in Lagrangian staggered-grid hydrodynamic (SGH) calculations for many years. Early examples of contact methods are discussed in Wilkins [37] and Cherry et al. [7]. Hallquist et al. [17] provides an overview of multiple contact algorithms used in various Lagrangian SGH codes dating back to HEMP [37]. Of particular interest, Hallquist et al. [17] describes the contact surface scheme used in TOODY [31] and later implemented in DYNA2D [36]. The contact method of TOODY uses a master–slave approach. The goal of this approach is to treat the nodes on the contact surface in a manner similar to an internal node."*

Keyphrases: *Contact methods, Lagrangian SGH, contact surface scheme, master–slave approach, TOODY*

ScienceIE dataset is applied in evaluating Chinese to English cross-lingual keyphrase extraction pipelines. To simulate the use case of extracting keyphrases in English out of Chinese scientific documents, the document set needs to be translated into Chinese before the experiment. However, within the available resource of this paper, manually translating the dataset is not practical. Thus, we use *Youdao* translator to conduct the dataset translation process. Following demonstrates a Chinese translation done by *Youdao* translator.

Document Translation: 多年来，在拉格朗日交错网格水动力*(SGH)*计算中，已经开发和使用了接触方法。在*Wilkins[37]*和*Cherry*等*[7]*中讨论了接触方法的早期例子。*Hallquist*等*[17]*提供了多种联系算法的概述，用于各种拉格朗日的*SGH*编码，可追溯到*HEMP[37]*。特别有趣的是，*Hallquist*等*[17]*描述了*TOODY[31]*中使用的接触面方案，后来在*DYNA2D*中实现*[36]*。*TOODY*的接触方法采用主从方法。这种方法的目标是用类似于内部节点的方式来处理接触面上的节点。

Keyphrase Translation: 接触方法, 拉格朗日*SGH*, 接触表面方案, 主从方式, *TOODY*

As presented in the translation example, the personal names and the special terms are well retained in the translation, and the expression is reasonably smooth. However, some errors still creep into the translated dataset despite the sufficient machine translation quality. For example, "hydrodynamic"should be translated as "流体力学"instead of "水动力". The original ScienceIE dataset and the Chinese translation version of the dataset are used for training the keyphrase extraction models and evaluation the cross-lingual keyphrase extraction pipelines.

Authors manuscript

The Weibo Named Entity Dataset [23] is proposed by He and Sun for training and testing Chinese named entity recognition models. This dataset consists of social media posts and comments, which are crawled from the Chinese social network Weibo. To test the performance of cross-lingual named entity recognition pipelines, we translate the Chinese named entity reference into English. Following is a social media post in the Weibo named entity dataset, two variations of corresponding machine translation results and two types of named entity references:

Chinese Social Media Post:
一节课的时间真心感动了李开复
English Translation 1:
The time of a lesson really touched Li Kaifu
English Translation 2:
The time of a class really moved Lee Kai-fu
Reference 1:
(This: O), (lesson: O), (really: O), (moved: O), (Li: PER), (Kaifu: PER)
Reference 2:
(Li: PER), (Kaifu: PER)

As we can see in the last example, a social media post can be translated differently, because of the differences between machine translation models. Therefore, full sentence reference is not applicable for evaluating the cross-lingual named entity recognition due to the mismatching of non-named-entity-words between machine translation and correct reference translation. However, the non-named-entity-words do not influence the evaluation result because in the reference, these words are considered as True-Negatives, which take no part in the calculation of Precision, Recall and F1 scores. Thus, for the evaluation of cross-lingual named entity recognition task, instead of translating the full sentences, we only translate the named entities in the reference.

B. Results

In this section, we present the evaluation results of the cross-lingual keyphrase extraction pipelines and cross-lingual named entity recognition pipelines. The comparisons are conducted in machine translation systems, keyphrase extraction algorithms, and processing sequences. We answer the research questions with the findings in the experiment and come up with hypotheses for possible reasons for the experiment outcomes.

Machine Translation

We first evaluate the translation quality of the machine translation systems with BLEU score [24], as BLEU score are commonly adopted for evaluating machine translation systems. Table I presents the 1-4gram BLEU scores of the five machine translation systems. Among the four commercial machine translation systems, *Google* translator is leading the four commercial machine translation systems on the BLEU scores with a relatively large margin while *Bing* has a slight edge over *Baidu*. The BLEU scores of *Youdao* translator are the lowest.

TABLE I: BLEU SCORES OF MACHINE TRANSLATION SYSTEMS

Metrics	*Youdao*	*Google*	*Baidu*	*Bing*
BLEU-1gram	0.5918	0.6380	0.5945	0.6002
BLEU-2gram	0.4297	0.4888	0.4377	0.4400
BLEU-3gram	0.3226	0.3829	0.3274	0.3110
BLEU-4gram	0.2461	0.3024	0.2491	0.2527

TABLE II: MONOLINGUAL KEYPHRASE EXTRACTION RESULTS

Algorithm	Precision	Recall	F1
Kea	0.1820	0.2307	0.2035
MultipartiteRank	0.2780	0.3385	0.3053
TF-IDF	0.1510	0.1910	0.1687

We can see that the translation candidates from *Youdao*, *Google*, *Baidu*, and *Bing* are somewhat similar. These four translation candidates correctly translate most of the key information in the document. Overall the performances of the commercial machine translation systems are sufficient. However, there still exist some subtle errors such as mistranslating the "combined clue"as "combination cues"or "combinatorial cues"in the translation, which may influence the performance of the keyphrase extraction and the named entity recognition processes. Hence, we can expect similar result in the experiment as well.

Keyphrase Extraction

To set a reference line for evaluating the cross-lingual keyphrase extraction, we test Kea, MultipartiteRank and TF-IDF algorithms on ScienceIE dataset to see the performance difference between the algorithms in the monolingual scenario. The scores of the three algorithms are presented in Table II:

From Table II it is clear that MultipartiteRank has significantly higher scores than Kea and TF-IDF. Kea achieves reasonably better scores than TF-IDF, which is expected since TF-IDF is often used as a baseline algorithm for keyphrase extraction task.

On the cross-lingual front, Figure 4 illustrates a comparative overview of scores of the cross-lingual keyphrase extraction pipelines. The cross-lingual keyphrase extraction pipelines are notated with four descriptions, for example, "kw-mt-bing-kea"where first two attributes ("mt-kw", "kw-mt") stand for the processing sequence of the pipeline; "mt-kw"means machine-translation-first processing sequence and "kw-mt"means machine-translation-later processing sequence. The third part ("*youdao*", "*google*", "*baidu*", "*bing*") of the description indicates the machine translation system. Last part ("kea", "mr", "tfidf") of the pipeline name refers to the keyphrase extraction algorithm in which "mr"is the abbreviation of MultipartiteRank.

Looking into the Figure , the pipeline "mt-kw-google-mr"achieves the best F1 score. The pipelines on the right of the figure have significantly higher scores than those on the left, which shows that translating the document is a better approach

Author manuscript

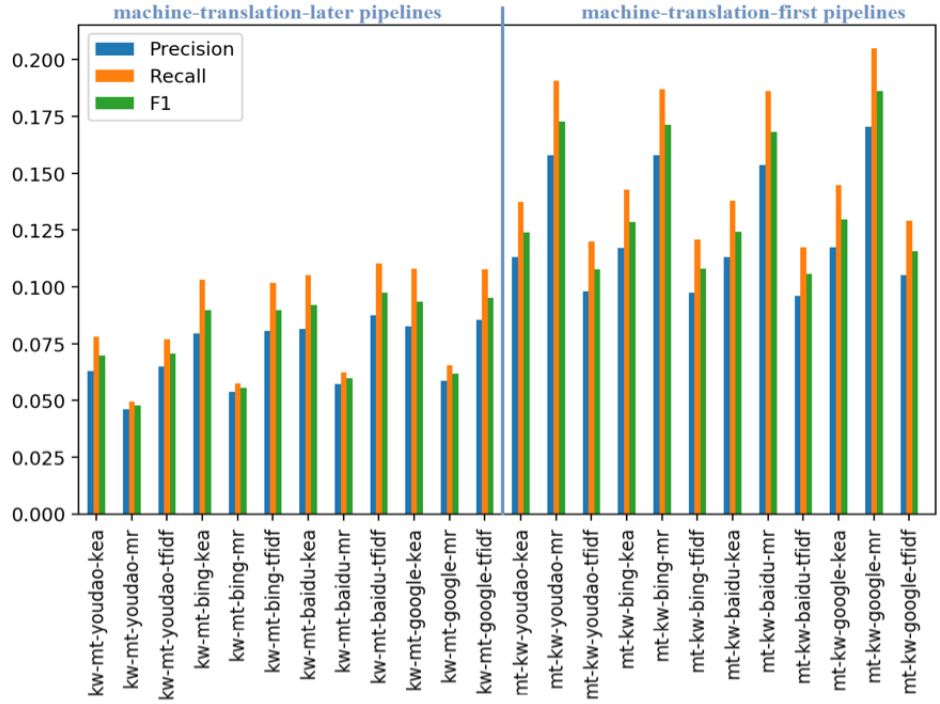

Fig. 4. Overall performance of cross-lingual keyphrase extraction pipeline based on scienceIE dataset

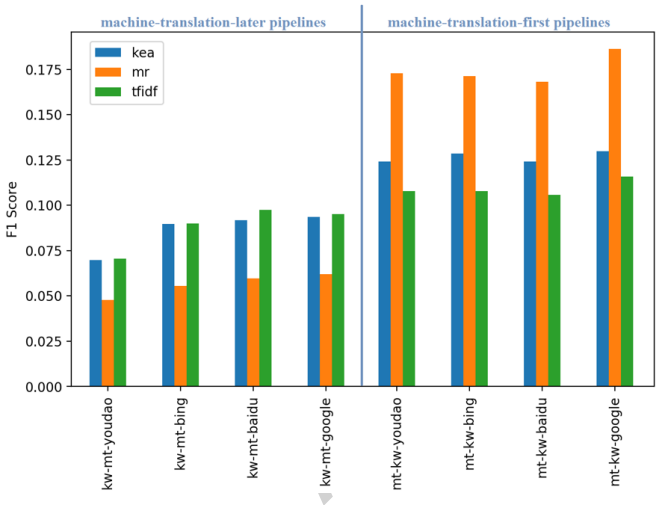

Fig. 5. Comparison of pipelines with different keyphrase extraction algorithms

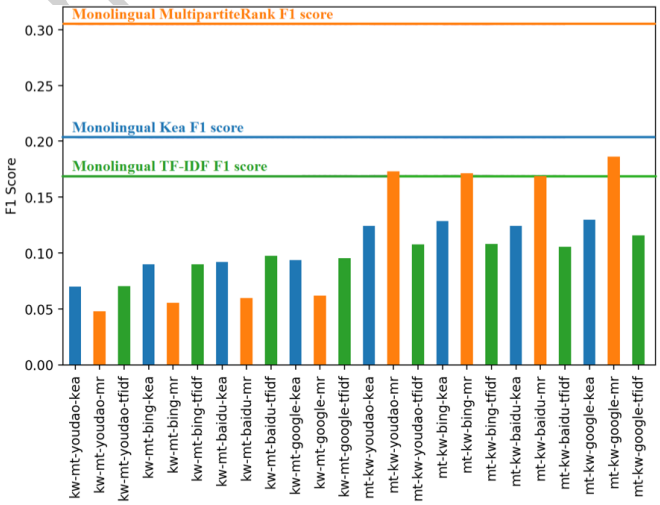

Fig. 6. Cross-lingual keyphrase extraction performance and the monolingual reference

than translating the extracted keyphrases. As can be seen in Figure 5, the processing sequences have particularly large impact on MultipartiteRank algorithm. MultipartiteRank performs much better on the machine-translation-first pipelines. We infer that the reason behind this phenomenon is likely to be the processing language of the algorithms, for different processing sequences imply different document languages for the keyphrase extraction algorithms to process. In this paper, the three algorithms extract English keyphrase in machine-translation-first pipelines and extract Chinese in machine-translation-later pipelines. Comparing to English, Chinese language is a relatively tricky language for the keyphrase extraction task due to its linguistic characteristics such as no marked word boundaries, which may also explain why the three algorithms are influenced differently by the the processing sequences. For example, Kea and TF-IDF are less affected in comparison to MultipartiteRank because Kea and TF-IDF use straightforward statistical information of the document such as term frequency and term position to predict keyphrases while MultipartiteRank analyzes more sophisticated features

TABLE III: MONOLINGUAL NAMED ENTITY RECOGNITION RESULTS

Algorithm	Precision	Recall	F1
Location	0.6795	0.5048	0.5792
Organization	0.7000	0.4298	0.5326
Person	0.7029	0.2687	0.3888
Total	0.6958	0.3431	0.4596

such as POS and phrase topics. Compared to simple statistical features, extracting the more complex linguistic features may bring more errors into the keyphrase extraction when processing Chinese documents. Incorporating with the arguments mentioned above, we deduce that the language difference is the reason why the processing sequence dramatically affects the performance of keyphrase extraction algorithms.

Comparing the pipelines with different machine translation systems in Figure 4 and Figure 5, we find that the score difference is inconspicuous. Overall, the performance of machine translation systems in the cross-lingual keyphrase extraction pipelines is comparable to the monolingual evaluation result with BLEU metrics. The *Google* translator is leading while the performances of *Baidu*, *Bing*, and *Youdao* translators are close. In general, the four commercial machine translation systems performs roughly at the same level.

To evaluate the disparity of the keyphrase extraction performance between monolingual and cross-lingual scenarios, we plot the F1 scores of each keyphrase extraction algorithm and its monolingual F1 score reference in Figure 6. We can see there exist significant disparities between cross-lingual keyphrase extraction pipelines and the corresponding monolingual references. The best cross-lingual F1 score of each algorithm is around 30% to 40% lower than their monolingual reference. However, from Figure 6, we observe that three cross-lingual pipelines ("mt-kw-google-mr", "mt-kw-youdao-mr", and "mt-kw-bing-mr") surpass the monolingual reference of TF-IDF. The pipeline "mt-kw-google-mr"achieves the best F1 score which is 10.4% higher than monolingual TF-IDF reference. For TF-IDF is a widely used baseline reference in keyphrase extraction task, surpassing it indicates that some cross-lingual keyphrase extraction pipelines with a better algorithm like MultipartiteRank are capable of extracting keyphrase with a quality comparable to the monolingual baseline performance.

Named Entity Recognition

To evaluate the performance of the Stanford named entity recognizer [22] as a monolingual reference, we test it with the Weibo named entity dataset. Table III shows the named entity recognition performance on the test set with separate evaluation of Location, Organization, and Person names:

As is presented in Table III, the recognizer performs better at recognizing location and organization names than person names in this test set. This is probably because that the test set is collected from social network Weibo, and thus lots of person names mentioned in the test set are nicknames

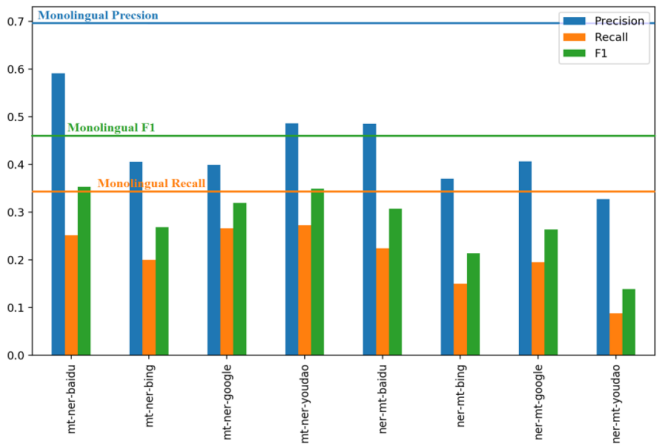

Fig. 7. Overall Performance of named entity recognition pipeline based on weibo named entity dataset

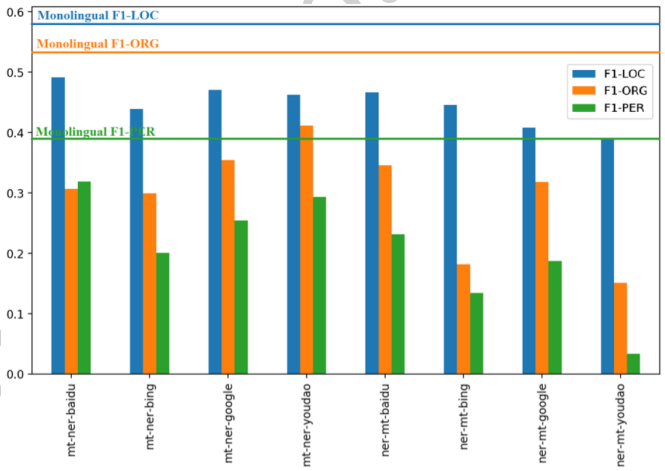

Fig. 8. Comparison of performance in recognizing different named entities

which may be difficult for the named entity recognizer to recognize. Regarding the evaluation results of the eight cross-lingual named entity recognition pipelines, Figure 7 illustrates these evaluation results with the corresponding monolingual references. The named entity recognition pipelines are notated with three descriptions, for example, "mt-ner-baidu", where first two parts ("ner-mt", "mt-ner") stand for the processing sequence of the pipeline; "mt-ner"refers to machine-translation-first pipelines while "ner-mt"refers to machine-translation-later pipelines; the third part of the description is the machine translation system ("google", "baidu", "bing", "youdao") implemented in the pipeline.

In the eight pipelines, "mt-ner-baidu"achieves the highest Precision and F1 score of 0.5911 and 0.3531 which are respectively 15% and 23% lower than the monolingual reference. The pipeline "mt-ner-youdao"has the best Recall of 0.2724 which is approximately 21% lower than the monolingual Recall score.

Regarding the processing sequence, we find that the

Authors manuscript

pipelines with machine-translation-first processing sequence achieve better performance than those with machine-translation-later processing sequence in recognizing Location, Organization and Person names. The machine-translation-first pipelines have on average around 20% higher F1 scores than the machine-translation-later pipelines. Using the F1 score as a single value evaluation metrics, Figure 8 provides a direct comparison of the performances of the cross-lingual named entity recognition pipelines in extraction location, organization and person names. On the whole, the cross-lingual named entity recognition pipelines perform the best in extracting Location names and perform poorest with Person names, which is similar to the monolingual reference. The only exception is the pipeline "mt-ner-baidu", for which the most difficult named entity to recognize is the Organization names.

We manually inspect the location names recognition result and notice that common location names such as "北京"(Beijing), "日本"(Japan) and "普罗旺斯"(Provence) are recognized accurately. There are some infrequent Location names like "万荣"(Wanrong) which is correctly translated but is recognized as a regular word. However, some special location names like restaurant names are neither translated nor recognized correctly. Restaurant "外婆家"is not translated correctly by all four machine translation systems. It is neither recognized as a location name in the original Chinese text nor in the translated English text. We suspect that these made up names are difficult both for machine translation systems and the named entity recognizer.

We also find a lot of unrecognized organization names do not have proper translations. For example, "虹口龙之梦"is a typical organization name which does not own a proper English name, for it is a local niche brand. It is translated to "Hongkou Dragon Dream"by *Google*, *Baidu* and *Bing* translators. With this translation, only the "Hongkou"part is recognized as an organization name.

Numerous of nicknames in the test set are not correctly translated and thus are not recognized as person names. The Chinese name "赵小晚"(Zhao Xiaowan) is mistranslated to "Zhao Xiao late". However, "赵小晚"is correctly recognized as person name in the original document, and "Zhao"of "Zhao little late"in the translated document is correctly recognized as person name as well. Hence, machine translation quality is likely to be a bottleneck of cross-lingual pipelines in recognizing person names.

V. Conclusion

In this paper, we implemented cross-lingual text processing pipelines to facilitate the automatic information extraction in cross-lingual scenarios. The main idea is to extract keyphrases and named entities in English from Chinese document by combining machine translation systems with the natural language processing techniques. We separately evaluated and compared the cross-lingual keyphrase extraction pipelines and cross-lingual named entity recognition pipelines with different combinations of components and sequences. Three keyphrase extraction algorithms were evaluated in this paper, namely Kea, MultipartiteRank, and TF-IDF. MultipartiteRank achieved the highest precision, recall and F1 scores both in monolingual and cross-lingual keyphrase extraction tasks. We tested four commercial machine translation systems in the cross-lingual text processing pipelines, which were *Baidu*, *Bing*, *Google* and *Youdao*. The results indicate that at the time of our evaluation the *Google* translator is the best option for cross-lingual keyphrase extraction task, as the cross-lingual keyphrase extraction pipelines with *Google* translator achieved the highest average Precision, Recall, and F1 score.

In the named entity recognition task, the *Baidu* translator backed pipelines achieved the best average Precision, Recall, and F1 score. Hence, the *Baidu* translator is the most recommendable for the cross-lingual named entity recognition pipelines. However, as all of these companies constantly work on improving their commercial machine translation systems, these results are likely to change over time and should be repeated to asses the different situation in the future. Thus, the evaluation results only represent the current state of these machine translation systems. We described two processing sequences for cross-lingual text processing, namely processing document before machine translation and processing machine translated document. The latter had significant advantages in both cross-lingual keyphrase extraction and named entity recognition pipelines. However, the evaluation results show that the performance disparity between monolingual and cross-lingual text processing is considerably wide.

Based on the evaluation results, we will focus our future work on trying to close the performance gap between monolingual and cross-lingual text processing. Currently, each configuration of the cross-lingual text processing pipeline only applies one algorithm, one machine translation system, and one processing sequence. Empirically, combining different models can enable better performance than each single model. Hence, in the future, we will utilize the approach of combining algorithms, translation systems or processing sequence (hybrid approach) to attempt to boost the performance of the cross-lingual text processing pipelines. To squeeze out better performance out of the same model combination, we plan to explore a variety of method to combine the models, for instance, using various machine learning approaches to select keyphrase candidates from the output of several keyphrase extraction algorithms or to merge several machine translation candidates into a possibly better translation. As combining models often implies extra cost such processing time, computation power and API fee, we will include an analysis for marginal performance gain, if existing, in the evaluation of the hybrid approach in cross-lingual text processing pipelines. Another direction for future work will be implementing more natural language processing techniques, such as Automatic Summarization and Topic Recognition, as well as more compatible languages into the cross-lingual text processing pipeline. The goal is to enable the multilingual and multifunctional text processing with a comparable performance of monolingual text processing. In addition to the functional enhancement of the cross-lingual text processing pipelines, we plan to

Authors manuscript

construct a dataset or evaluation metrics for cross-lingual text processing pipelines, so that we can objectively evaluate the pipelines without the bias in the translation preprocessing.

ACKNOWLEDGMENT

This work is part of the project '"ELLI 2 - Excellent Teaching and Learning in Engineering Sciences" and was funded by the Federal Ministry of Education and Research (BMBF), Germany.

REFERENCES

[1] Mccarley and J. Scott, "Should we translate the documents or the queries in cross-language information retrieval?" in *Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999.

[2] C. Aone, H. Blejer, M. E. Okurowski, and C. Van Ess-Dykema, "A hybrid approach to multilingual text processing: Information extraction and machine translation," in *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, 1994.

[3] D. A. Hull and G. Grefenstette, "Querying across languages: a dictionary-based approach to multilingual information retrieval," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 49–57.

[4] X. Saralegi and M. L. De Lacalle, "Dictionary and monolingual corpus-based query translation for basque-english clir." in *LREC*, 2010.

[5] D. Wu and D. He, "A study of query translation using google machine translation system," in *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*. IEEE, 2010, pp. 1–4.

[6] F. Ture, J. Lin, and D. Oard, "Combining statistical translation techniques for cross-language information retrieval," *Proceedings of COLING 2012*, pp. 2685–2702, 2012.

[7] M. Basaldella, M. Helmy, E. Antolli, M. H. Popescu, G. Serra, and C. Tasso, "Exploiting and evaluating a supervised, multilanguage keyphrase extraction pipeline for under-resourced languages," in *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*. Incoma Ltd. Shoumen, Bulgaria, 2017, pp. 78–85.

[8] D. W. Oard and P. Hackett, "Document translation for cross-language text retrieval at the university of maryland," in *TREC*. Citeseer, 1997, pp. 687–696.

[9] K. Sudo, S. Sekine, and R. Grishman, "Cross-lingual information extraction system evaluation," in *Proceedings of the 20th international Conference on Computational Linguistics*. Association for Computational Linguistics, 2004, p. 882.

[10] K. Parton, K. R. McKeown, B. Coyne, M. T. Diab, R. Grishman, D. Hakkani-Tür, M. Harper, H. Ji, W. Y. Ma, A. Meyers *et al.*, "Who, what, when, where, why?: comparing multiple approaches to the cross-lingual 5w task," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 423–431.

[11] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, "Word translation without parallel data," in *International Conference on Learning Representations*, 2018.

[12] M. Al-Suhaiqi, M. A. S. Hazaa, and M. Albared, "Arabic english cross-lingual plagiarism detection based on keyphrases extraction, monolingual and machine learning approach," 2019, pp. 1–12.

[13] K. Parton, A. Meyers, S. Stolbach, A. Sun, G. Tur, X. Wei, S. Yaman, K. R. Mckeown, B. Coyne, and M. T. Diab, "Who, what, when, where, why," *Nature Nanotechnology*, vol. 3, no. 4, p. 179, 2008.

[14] J.-M. Torres-Moreno, "Multi and cross-lingual summarization," *Automatic Text Summarization*, pp. 151–177, 2014.

[15] C.-Y. Lin, "Machine translation for information access across the language barrier: the must system," in *Machine Translation Summit VII*, 1999, pp. 13–17.

[16] D. K. Evans, J. L. Klavans, and K. R. McKeown, "Columbia newsblaster: Multilingual news summarization on the web," in *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics, 2004, pp. 1–4.

[17] T. Huang, L. Lei, and Y. Zhang, "Multilingual multi-document summarization with enhanced hlda features," in *China National Conference on Chinese Computational Linguistics*, 2016.

[18] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 01, pp. 157–169, 2004.

[19] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "Kea: Practical automated keyphrase extraction," in *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI Global, 2005, pp. 129–152.

[20] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," *arXiv preprint arXiv:1803.08721*, 2018.

[21] F. Boudin, "pke: an open source python-based keyphrase extraction toolkit," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 2016, pp. 69–73.

[22] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[23] H. He and X. Sun, "F-score driven max margin neural network for named entity recognition in chinese social media," *arXiv preprint arXiv:1611.04234*, 2016.

[24] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation." in *Proc Meeting of the Association for Computational Linguistics*, 2002.

Authors manuscript