Cross-Lingual Sentence Extraction for Information Distillation

Adish Kumar Singla[1,2], *Dilek Hakkani-Tür*[1]

[1]ICSI, Berkeley, CA
[2]EPFL, Switzerland
adish.singla@epfl.ch, dilek@icsi.berkeley.edu

Abstract

Information distillation aims to analyze and interpret large volumes of speech and text archives in multiple languages and produce structured information of interest to the user. In this work, we investigate cross-lingual information distillation, where non-English (source language) documents are searched for user queries that are in English (target language). We propose to perform distillation both on the original source language data and their English translations output by machine translation, and combine the two outputs. We experimentally show that combination approach results in 8% to 16% absolute (13% to 31% relative) F-measure improvement over the previous work.

Index Terms: information distillation, sentence extraction, cross-lingual processing, and classification model combination.

1. Introduction

As the amount of information available to users increase, accessing information in an efficient way becomes difficult. The goal of information distillation is to extract useful pieces of information from massive multi-lingual audio and textual document sources given a user's query. In the framework of the DARPA GALE project, these sources are English, Arabic and Mandarin newswire, blogs, broadcast news and conversations. The participants are given a set of query templates with variable slots, sample queries and answer keys (a set of relevant sentences for each query) in advance. At run-time, participating systems are expected to find answers for new query examples that have a form of one of the query templates. In mono-lingual information distillation, both the user's query and documents to be searched are in same language. In cross-lingual distillation, the user query is in one language (usually called the target language), and documents are in other languages (called source language). In such a case, one can either search the query in source language documents and then translate the output to the target language (in this case, English) or search the machine translation (MT) output in target language (in this case, Arabic or Mandarin), or search both and combine the outputs.

Our distillation approach is based on using document retrieval for finding relevant documents in response to a given query, and statistical classification to extract sentences as snippets from the relevant documents [1]. The set of selected sentences is then reduced by finding and eliminating redundancies. To train the sentence extraction models, negative and positive (relevant) examples are extracted using the given answer keys, which have the relevant snippets and the corresponding document identifiers for each query. Each sentence and query is then processed to extract a set of syntactic and semantic features represented in a graph structure [2].

While source language answer keys can be used to train sentence extraction models from source language documents, on the English side, both the English answer keys and machine translation of the source language answer keys and corresponding documents can be used to train sentence extraction models. The noise introduced by automatic translation can reduce the accuracy of the distillation process for the second approach. However, there is much more annotated data available from the English sources.

The first goal of this paper is to check if the current, word-based distillation approach is suitable for Mandarin and Arabic, and to compare information distillation systems that can be used before (for the source language) and after machine translation (for the target language). For example, in Mandarin, words are not separated by space, and errors in automatic word segmentation can cause problems for a classifier using word n-grams from the source language as features.

The second goal of this paper is to see if the combination of the source and target language information distillation is useful. We experiment with several ways of combining source and target language models to benefit from the absence of automatic translation errors on the source language sentences and abundance of English data. We experimentally show that combination approach results in 8% absolute F-measure improvement for Mandarin[1] and 16% absolute F-measure improvement for Arabic[2] over the previous work.

Section 2 summarizes related work for information distillation and cross-lingual processing. Section 3 describes our approach for mono-lingual and cross-lingual distillation. Section 4 presents experimental results and discussion.

2. Related Work

For information distillation in the framework of the DARPA GALE project, Schiffman *et al.* [3] used an approach based on information extraction (IE) and retrieval (IR). In the first pass, their system requests only high precision documents from IR. Then, IE relations and events found in the returned documents are used to select relevant sentences. Finally, words from these sentences are used to augment the original IR query, which is used to retrieve a new set of documents. In our previous work, we used IE annotations to filter out documents retrieved by IR [4] and then extended this work for cross-lingual document retrieval by combining IE annotations in source and target languages [5]. For questions relying on free-text topic formulations, Levit *et al.* [6] used the deep semantic representations of a question and a candidate answer, that is based on the extracted predicate-argument structures embedded in an error-tolerant instantiation mechanism. The resulting instantiation score is used to determine the appropriateness of the answer.

[1]13% relative, from 0.62 to 0.70
[2]31% relative, from 0.52 to 0.68

Recently, there have been a number of studies on cross-lingual question answering (QA) and information retrieval in the text retrieval conferences (TREC) [7] and workshops held at the meetings of Cross Language Evaluation Forum (CLEF) [8]. Xu *et al.* [9] present cross lingual information retrieval of Arabic documents from English queries. They discuss number of language processing techniques that can be applied on the Arabic side because of the different linguistic characteristics of the Arabic language as compared to English. Then they extend this work with various other language processing techniques without much improvement in the overall IR accuracy [10]. Kwok *et al.* [11] present a QA system in a cross-lingual framework. Their system returns about 50 bytes answers for English queries based on key-word spotting and pattern matching. For Arabic documents, translations of original queries were used. Bos and Nissim [12] discuss cross-lingual QA systems where answers are extracted from documents in the same language as that of the query and then the answers are translated to the targeted language, in contrast to other systems which translate the queries to the target language instead, as in the other CLEF publications [13].

3. Approach

Our work is based on statistical sentence extraction approach proposed in [1] for mono-lingual processing of English data as described in the next subsection.

3.1. Mono-Lingual Information Distillation

To train sentence extraction models, positive and negatively marked sentences are taken from labeled data. Given a set of queries q_i for a template Q_T, and a set of sentences marked as relevant to q_i in all the relevant documents, training data S is formed:

$$S = \{(x_1, 1), (x_2, 1), \ldots, (x_n, 1), (y_1, 0), (y_2, 0), \ldots, (y_m, 0)\}$$

where $x_i, i = 1, \ldots, n$ are the relevant sentences and $y_j, j = 1, \ldots, m$ are the irrelevant sentences. A classification model is trained for each query template Q_T. This model estimates $P(c|s_k)$, $c \in \{0, 1\}$ for each sentence s_k in the documents returned by information retrieval. Sentences that have $P(c|s_i) > \tau_q$ are returned as relevant, where the threshold τ_q is estimated using a validation set for each query template Q_T.

3.2. Cross-Lingual Information Distillation

In the cross-lingual framework, information distillation can also be performed by using above mono-lingual framework, by building statistical models on target language side using the machine translation output of the source language (Arabic, Mandarin) documents. These models output a probability $P_{MT}(c|s_k)$, $c \in \{0, 1\}$ for each sentence s_k in the translated sentences of the candidate documents. In this case, English documents can also be used in addition to the MT output data during training, outputting the probability $P_{MT+ENG}(c|s_k)$. Usually, there are more natural language processing (NLP) resources and tools on English language side. However, the noise introduced by MT into the documents usually degrades the performance of the NLP tools.

On the other hand, another option is to build models directly on source language side, and estimating $P_{SRC}(c|s_k)$. Even though we have less data and less linguistic resources on the source language side, this data is free of noise introduced by MT errors.

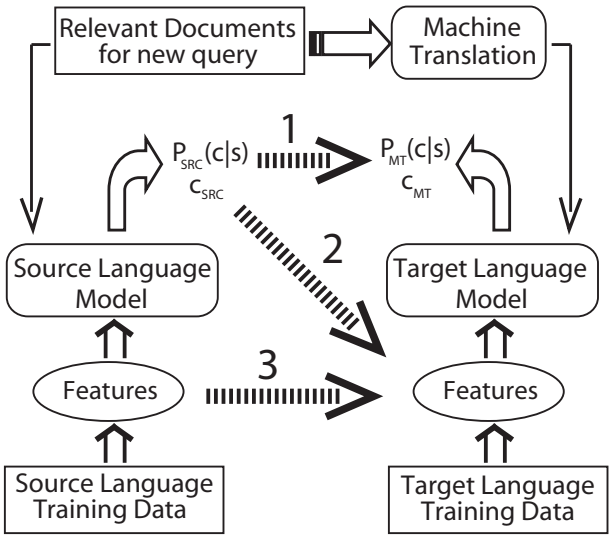

Figure 1: *Three ways of combining sentence extraction in source and target languages.*

Another approach, that we have followed in this work, is that we can perform distillation on both source and target language sides and then merge the information present in source language side (original documents) and in target language side (MT output documents). There are at least three ways of using information from both sides, as also shown in Figure 1:

1. **Posterior Probability Interpolation**: For each class c and sentence s, the posterior probability estimated by the two models, $P_{SRC}(c|s)$ from the source language model and $P_{MT}(c|s)$ from the target language model can be interpolated, estimating the new probability:

   $$P_{FINAL}(c|s) = \lambda \times P_{MT}(c|s) + (1 - \lambda) \times P_{SRC}(c|s)$$

   where $\lambda \in [0, 1]$ is estimated from a held-out data set.

2. **Using the output from one model as feature in the other:** The probability estimate, $P_{SRC}(c|s)$, and/or the class estimate $\hat{c}_{SRC}$ from the source language model can be used as a feature in the target language model (or vice versa).

3. **Combining features from Source and Target Language Sentences and MT output:** Instead of training two separate models from source and target language data, the set of features computed from the two data sets can be combined and a single model is learned from them. Assume S_i are the the set of features extracted from the sentences in the source language, and T_i are the set of features extracted from the machine translation output corresponding to S_i, and c_i are the relevance annotations (where $i = 1, ..., N$, and N is the number of examples from source language documents). Then one can learn a model from $(S_i \cup T_i, c_i)$, instead of the two models learned from (S_i, c_i) and (T_i, c_i). In order to utilize the answer keys from English documents, with features E_j (where $j = 1, ..., M$, and M is the number of examples from English documents), and relevancy annotations e_i, one can treat English examples in the same framework by considering unknown values for source language features. The classification algorithm

used should be able to deal with unknown feature values for this case.

4. Experiments and Results

Based on the framework of [1], we assume that IR engine returns a set of relevant documents for the queries on which we apply our sentence extraction approach. The corpus that we use is a collection of Arabic, English and Mandarin broadcast news, broadcast conversations, newswire and other written news forms such as blogs, and so on as described in the next subsection.

4.1. Data sets and Evaluation metric

We use query template 16 from the GALE project year 1 and year 2 distillation data sets, since there is annotated data from both of the non-English data sources for this template. Template 16 has the following form:

`Describe attacks in` **location** `giving location (as specific as possible), date and number of dead and injured.`

Here **location** is a variable slot specified for every query example.

For GALE year 1 data set, there are a total of 24 queries for template 16, all 24 have answer keys from English documents, 11 of them have answer keys from Arabic documents and 12 of them have answer keys from Mandarin documents. For GALE year 2 data set, there are 9 queries with answer keys for English, 8 for Arabic and 9 for Mandarin. Table 1 shows properties of these data sets.

We use BoosTexter [14] for sentence classification, with n-fold cross-validation with total of 19 rounds for Mandarin data and 21 rounds for Arabic data. 20% of the training data is used as validation set while computing the optimal number of iterations and probability threshold values, while data from one query is the test set for each round. Note that, BoosTexter assigns a probability of being relevant to each of the sentence and can deal with unknown valued features. In order to compare different systems, we use macro-averaged F-measure over query template examples.

The Arabic and Mandarin documents were translated into English using SRI's machine translation system [15].

Language	Number of			
	Queries	Docs	Relevant Sentences	Irrel. Sentences
Arabic	19	292	480	1,344
English	33	1,239	3,130	12,799
Mandarin	21	493	1,085	2,693

Table 1: Data sets used in the experiments: number of queries, documents, relevant and irrelevant sentences for all three languages.

4.2. Results from mono-lingual experiments

In order to check the effect of automatic word segmentation for Mandarin, we first experimented with the initially distributed 12 query example subset of Mandarin data. This corresponds to the year 1 data set. We experimented with a publicly available word segmentation tool[3], as well as the word segmentation tool from University of Washington [16]. We compared the performance of word n-gram models with character n-gram models, and found that F-measure does not change significantly after 3-grams for words and 5-grams for characters. The corresponding F-measure results are listed in Table 2, and character n-grams are significantly better than word n-grams. So, in all the following experiments, we used character n-grams for Mandarin.

Features	F-measure
Words 3-grams (MandarinTools.com)	0.46
Word 3-grams (UW segmenter)	0.49
Character 5-grams	0.55

Table 2: Results for Mandarin data with various feature sets.

4.3. Results from cross-lingual experiments

Next set of experiments compare source language models with the corresponding models in English with machine translation output (MT), as well as other available data from English documents (ENG).

Train	Test	Arabic	Mandarin
Chance		0.42	0.47
SRC	SRC	0.60	0.60
MT	MT	0.58	0.64
ENG	MT	0.52	0.62
ENG+MT	MT	0.53	0.61

Table 3: F-measure results on Mandarin and Arabic data sets, when only the source or the target language data is used. SRC and MT means that source language text, machine translation output text was used in the experiments, respectively. ENG represents the training data for the query template from English documents.

Table 3 lists results from these experiments. Chance performance is the F-measure when all sentences in all relevant documents are listed as relevant. The recall is 100% for the chance performance, but precision is low. Note that, in the SRC and MT training data sets, the number of training examples are the same, whereas the ENG data set is much bigger. The test data set is the same for all these experiments, except, in the first experiment (SRC), features are extracted from the original source language sentences, whereas in all the rest of the experiments, features are extracted from the corresponding machine translation output. For Arabic, the results are better when source language data is used instead of machine translation output. In Mandarin, MT output results in better F-measure. This is probably due to the more meaningful and longer span features from word n-grams that are used in the English experiments. Our previous system in the project evaluations was using ENG data as the training set, and the MT data as the test set. Apparently this was sub-optimal.

Table 4 has the results from the experiments, where source and target language information is used in the experiments. In the first experiment, the probabilities from the source and target language models are interpolated. The target language model that results in the best performance on the held-out data set is

[3] Available from `http://www.mandarin.tools.com`.

chosen for experiments on the test set. The second and third experiments include the output from the source language experiment as a feature in the target language experiment. In the first of these, the probability output from the source language classifier is used, in the second one, the relevancy decision is used. The fourth experiment uses only data from the source language, but also includes features from MT output sentences. The fifth one is similar to the fourth, but the training data is much larger, since the data from English sources is also used.

Method	Arabic	Mandarin
1	**0.68**	**0.70**
2 with P_{SRC}	0.58	0.67
2 with $\hat{c}_{SRC}$	0.62	0.63
3 (MT & SRC)	0.56	0.62
3 (MT+ENG & SRC)	0.58	0.65

Table 4: Cross-lingual distillation F-measure results, when two information sources are combined. As in Table 3, MT represents the automatic translations of non-English data, ENG represents data from English documents.

For both languages, one or more of the combination methods helped improve performance. The best results are obtained using linear interpolation for both. These results are significantly better than the results in Table 3 and the results with other combination methods.

5. Conclusions

This work presents a data driven approach for sentence extraction successfully employed in the cross-lingual framework. 8% to 16% absolute performance gains have been shown by successfully using the information from the source language side along with MT data as compared to using only English data sources. The first combination approach of building two separate models and interpolating their probability outputs has given the best results for both Arabic and Mandarin. For other combination approaches in which we add source language side features to target language side models would probably perform better with sufficient amount of training and validation data.

In this study, we only used n-grams lexical features and not any other semantic features such as named entities, resolved coreferences and so on. We did not perform any preprocessing on the documents i.e. no stop word removal, no stemming etc. We have demonstrated the affect of Mandarin word segmentation on distillation performance and we feel that by applying number of preprocessing steps on source language side, we can gain better performance.

We plan to extend this work by also adding semantic features. We also plan to use the knowledge of slots filled in by user in the query template as features as demonstrated in [2]. However, we have only limited linguistic resources on Arabic and Mandarin and hence it may not be possible to directly employ the advanced framework of [2] in our work.

Acknowledgments: We thank Jing Zheng for providing us the machine translation output used in the experiments, and Gokhan Tur and Michael Levit for many helpful discussions. This work is partly supported by the Swiss National Science Foundation through the research network IM2 and the Defense Advanced Research Projects Agency (DARPA) GALE project, under Contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

6. References

[1] D. Hakkani-Tür and G. Tur, "Statistical sentence extraction for information distillation," in *Proceedings of ICASSP*, Hawaii, April 2007.

[2] M. Levit, D. Hakkani-Tür, G. Tur, and D. Gillick, "Integrating several annotation layers for statistical information distillation," in *Proceedings of ASRU*, Japan, December 2007.

[3] B. Schiffman, K. McKeown, R. Grishman, and J. Allan, "Question answering using integrated information retrieval and information extraction," in *Proceedings of HLT/NAACL*, Rochester, April 2007.

[4] D. Hakkani-Tür, G. Tur, and M. Levit, "Exploiting information retrieval annotations for document retrieval in distillation tasks," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007.

[5] D. Hakkani-Tür, H. Ji, and R. Grishman, "Using information extraction to improve cross-lingual document retrieval," in *Proceedings of RANLP Workshop on Multi-source Multilingual Information Extraction and Summarization*, Bulgaria, 2007.

[6] M. Levit, E. Boschee, and M. Freedman, "Selecting on-topic sentences from natural language corpora," in *Proceedings of the Interspeech 2007*, Antwerp, Belgium, 2007.

[7] F. Gey and D. Oard, "The trec-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries," in *Proceedings of TREC*, 2001.

[8] M. Kluck and F. Gey, "The domain-specific task of CLEF - specific evaluation strategies in cross-language information retrieval," in *Workshop of the Cross-Language Information Evaluation Forum, CLEF 2000*, Lisbon, Portugal, 2000.

[9] J. Xu, A. Fraser, and R. Weischedel, "Cross-lingual retrieval at BBN," in *Proceedings of TREC*, 2001.

[10] A. Fraser, J. Xu, and R. Weischedel, "Cross-lingual retrieval at BBN," in *Proceedings of TREC*, 2002.

[11] K. Kwok, L. Grunfeld, N. Dinstl, and M. Chan, "Question-answer, web and cross language experiments using pircs," in *Proceedings of TREC*, 2001.

[12] J. Bos and M. Nissim, "Cross-lingual question answering by answer translation," in *Working Notes of CLEF 2006*, 2006.

[13] C. Peters, "Results of the clef 2003 cross-language system evaluation campaign," in *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway, August 2003.

[14] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000. [Online]. Available: citeseer.ist.psu.edu/schapire00boostexter.html

[15] J. Zheng, W. Wang, and N. F. Ayan, "Development of SRI's translation system for broadcast news and broadcast conversations," in *Proceedings of Interspeech*, Brisbane, Australia, 2008.

[16] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozakil, "Investigation on mandarin broadcast news speech recognition," in *Proceedings of Interspeech*, Pittsburgh, PA, 2006.