Combining Evidence from Homologous Datasets

Ao Feng and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{aofeng, allan}@cs.umass.edu

ABSTRACT

With Machine Translation and/or Automatic Speech Recognition, there can be different versions of the same data with distinct expressions. We argue that combining evidence from these "homologous" datasets can give us better representation of the original data, and our experiments show that a model combining all sources outperforms each individual dataset in retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Query formulation

General Terms

Algorithms, Experimentation, Languages, Performance

Keywords

Evidence combination, CLIR, Machine translation, Indri

1. INTRODUCTION

Nowadays many information retrieval collections are not limited to a single language, like those in the Text REtrieval Conference (TREC), with data in other languages machine translated (MT) into English. For projects focused on news like Topic Detection and Tracking (TDT), data may also come from multiple media (newswire, web, radio, TV, etc.), and broadcast news often comes with transcripts from Automatic Speech Recognition (ASR). The variety of sources and MT/ASR systems results in different versions of the same dataset, and their quality varies. Extensive experiments are usually required to pick the "best" version that achieves the highest accuracy in information retrieval.

Without enough relevance judgment, it is difficult or sometimes impossible to decide which version is the "best". Fortunately, there is an alternative to that - if the quality of different versions cannot be compared directly, why not use all of them? It is safe to assume that all MT/ASR systems make mistakes, and the errors are usually not all the same. Therefore, a combination model that processes all different versions concurrently and merges their output is more likely to compensate for each other.

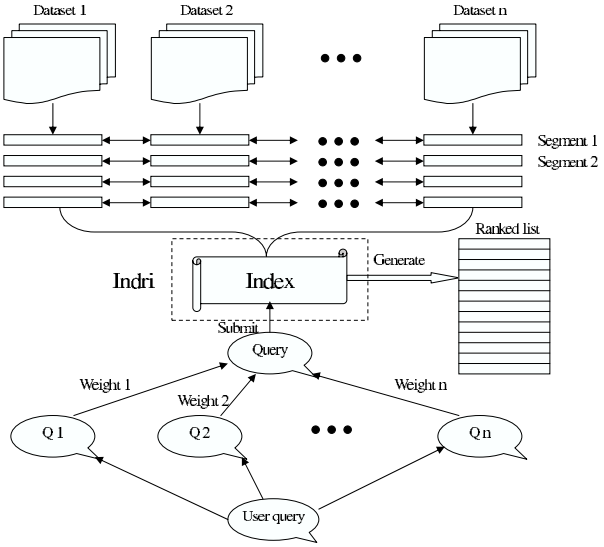

Figure 1: The combination model

It has been widely accepted in the Information Retrieval (IR) community that combining multiple sources of evidence can improve retrieval performance [3]. However, most experiments focus on different representations of the information need [1] or multiple document models [6], and we have not seen any attempt to utilize parallel data. Metasearch [5] merges the results of multiple algorithms, while our model combines the evidence from different content.

Cross-Language Information Retrieval (CLIR) finds documents in one language with queries in other languages. Commonly used methods in CLIR include MT, parallel corpora, query expansion, pseudo-relevance feedback, etc. [2]. The combination model is not restricted to one language, and we use both automatic and manual query translation in the experiments.

2. COMBINATION MODEL

Figure 1 shows the structure of the combination model. It is composed of two main steps - the indexing step and the

MLIA workshop at SIGIR'06, August 10, 2006, Seattle, Washington, USA.

<DOC>
<DOCNO> XIN20030701.1130.0132 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 20030701_11:53:00 </DATE_TIME>
<TEXT>
<MAN>
...
新华社伦敦 7 月 1 日电
温布尔登网球公开赛女单四分之一决赛 1 日举行，美国的威廉姆斯姐妹、比利时名将克利斯特和埃南击败各自的对手，均获得了半决赛权。
...
</MAN>
<MTa>
...
Xinhua News Agency, London, July 1 Wen Wimbledon women's singles one quarter finals held in 1st, the Williams sisters, Belgium Cleveland and Pierce defeated their respective opponents in the semi finals.
...
</MTa>
<MTb>
...
the Xinhua News Agency, London, July 1 electric wimbledon tennis finals single female one fourth of the 1st meeting of the United States, the Williams sisters from Belgium, Mr Ernest generals and Hainan defeat of their respective adversary, have gained the right and a half finals.
...
</MTb>
<MTc>
...
Xinhua News Agency, London, July 1 the - wenbuerdeng tennis public competition female list 1/4 plays in the finals on 1st to hold, US's weilianmusi sisters, Belgian great soldier Kills especially defeats the respective match with Egypt south, has obtained the semi-final power.
...
</MTc>
</TEXT>
</DOC>

Figure 2: A "document" that contains four different versions: one in Mandarin and others from three MT systems

query step.

The upper half in Figure 1 is the indexing step. In this figure, there are n parallel datasets. They can be in different languages or even different media, but here we only deal with text (due to the fact that we are lack of tools for processing multimedia data). Each corpus is then broken down into small segments. For text, these units are usually documents/news stories, but they can also be passages. Segments from different datasets must be correctly aligned so that they have corresponding content. In the next step, aligned segments from various sources are merged into a "document", marked by separate tags. Figure 2 shows a sample "document", in which tag MAN is for the source Mandarin data, while MTa, MTb and MTc are for three different MT systems.

All the merged data are then indexed by Indri, with "documents" as the basic units. Indri is an information retrieval system based on language modeling and inference network [4]. It has the ability to index individual fields. The n datasets have their own tags in the index (like MAN, MTa, MTb and MTc in Figure 2), so each version is separately accessible by Indri.

When the query is in one language and the document is in another, we cannot match them directly because their vocabularies do not have any overlapping. Under this condition, the query needs to be translated into the source language when necessary. Sometimes even corpora in the same language may use different words (note Williams vs. weilianmusi in Figure 2), which is hard to deal with unless we have the translation tables for all MT systems. Based on the confidence in different versions, there can be various weights associated with the individual fields. Here is the final query submitted to the Indri index:

$$\#weight(w_1\ q_1.tag_1\ w_2\ q_2.tag_2\ \ldots\ w_n\ q_n.tag_n) \qquad (1)$$

$\#weight$ is a belief operator in Indri. It has an even number of arguments, where the odd-indexed ones (w_i) are weights assigned to the query in the next argument. $q_i.tag_i$ means that query q_i (the original user query or some translation of it) is used to match the text in the i-th field identified by tag_i. The similarity score of a query q and a document D is calculated by:

$$sim(q, D) = \sum_{i=1}^{n} w_i sim(q_i, D_{tag_i}) \qquad (2)$$

where $sim(q_i, D_{tag_i})$ is the similarity score between q_i and the text in the tag_i field of D[1]. The retrieval result returned by query q is a ranked list of documents sorted by the similarity in Equation 2.

System	MAP	P-value$_{max}$
MTa (E)	0.4876	
MTb (E)	0.4443	
MTc (E)	0.3308	
$Man1$ (C)	0.2704	
$Man2$ (C)	0.4209	
$Combine1$ (E)	0.4752	0.6990
$Combine2$ (E+C)	0.5042	0.0978
$Combine3$ (E+C)	0.5004	0.1076
$Combine1_{opt}$ (E)	0.5014	0.0179*
$Combine2_{opt}$ (E+C)	0.5272	0.0067*
$Combine3_{opt}$ (E+C)	0.5159	0.0445*

Table 1: Ad-hoc retrieval results of individual and combined runs

3. EXPERIMENTS

It is possible that the datasets in the combination model are in different media, but we processed only text in our experiments. The reason is that we do not have the necessary tools/models to convert/retrieve multimedia data. The text collection contains only newswire, since different ASR systems for broadcast news usually have various document boundaries, which makes alignment between different versions quite difficult.

We used the TDT-5 collection in our experiments. It is a news corpus collected from English, Mandarin and Arabic sources, and the time spans from April to September 2003. This collection does not contain broadcast news data, only newswire sources. We used the Mandarin data from July, August and September, making up 602 files containing a total of 27,723 news stories (documents).

We obtained three machine translations of the 27,723 stories from three different systems. (None of them is the one included in the TDT-5 corpus.) The translations represent current state-of-the-art technologies, but we do not identify the systems here because they are preliminary work, calling them MTa, MTb and MTc respectively. MTa and MTb are both based on statistical models, while MTc uses rules.

From the 250 topics in TDT-5, we selected 35 that have at least 5 on-topic stories in our collection. The topic titles (in English as provided by the LDC) are used as queries.

All experiments are on the combined index, with different weight settings. The results are shown in Table 1.

[1]Details of the retrieval model can be found here: http://ciir.cs.umass.edu/~metzler/indriretmodel.html.

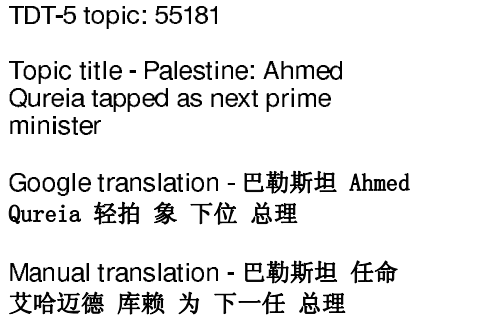

Figure 3: Comparison of query translation - automatic vs. manual

- MTa, MTb and MTc: Only one of the three MT versions gets weight 1, all other weights (including Mandarin) are set to 0. Topic titles are used without any translation.
- $Man1$: We translated the English topic titles into simplified Chinese with Google translation (http://www.google.com/language_tools), and then manually segmented terms. We use these queries to match the original Mandarin data (field MAN in Figure 2) and set its weight to 1. Each English source gets a weight of zero.
- $Man2$: The translation quality in $Man1$ is not very good. In fact, some terms (especially names) were not translated at all (see Figure 3). To get Mandarin queries in better quality, we manually translated the topic titles and replaced the queries in $Man1$ with the output. The retrieval accuracy is obviously improved with the manually translated queries.
- $Combine1$: All English versions (MTa, MTb and MTc) are combined with equal weights.
- $Combine2$: We combine all English versions and original Mandarin. The topic title is used as the English query, and the output from Google translation is the Mandarin query. Each of the four fields gets the same weight.
- $Combine3$: Same as $Combine2$ except that the Mandarin queries are manually translated.
- $Combine1_{opt}$, $Combine2_{opt}$ and $Combine3_{opt}$: The weights in the combined versions are tuned to maximize the Mean Average Precision (MAP). In $Combine1_{opt}$, (MTa, MTb, MTc) = (1.25, 1.0, 0.15). For $Combine2_{opt}$, (MTa, MTb, MTc, $Man1$) = (1.25, 1.0, 0.15, 0.7). (MTa, MTb, MTc, $Man2$) = (1.25, 1.0, 0.15, 0.85) in $Combine3_{opt}$.

The first data column in Table 1 shows the MAP of all 35 queries. The corresponding number in the second column (only for the combined runs) is the maximum of the P-values from the Wilcoxon rank-sum tests between the combined run and each composing version. For example,

$$Pvalue(Combine1) = max(Pvalue(Combine1, MTa), Pvalue(Combine1, MTb), Pvalue(Combine1, MTc)) \quad (3)$$

A number with * means that the combined run has significant improvement over all individual versions at 95% confidence level.

Without any parameter tuning, the combined model is at least comparable to the best individual version. The improvement is more obvious when we mix sources with different vocabularies. Note that $Man1$ has the lowest MAP (because of the translation quality), but incorporating it into the combination model improves the performance remarkably. With proper weight setting, the combination model can be significantly better than any individual version.

Another interesting observation is the comparison between $Combine2$ and $Combine3$, for both the non-optimized and optimized combination. Although the accuracy in $Man1$ is much lower than that in $Man2$, we do not see any improvement in the combined run when we replace the Google translated queries with the manual ones. The reason for that is still unclear to us, but we can draw a conclusion from the experiments - the performance of the combination model relies more on the heterogeneity than on the quality of individual sources.

4. CONCLUSIONS

Datasets often come in different versions. Instead of selecting the best of them with great difficulty, a combination model can be designed to utilize all available evidence concurrently. Our experiments show that this model yields better results than any of the individual versions, especially when the sources are in multiple languages. Due to the limitation of available tools, we do not have any cross-medium experiment. However, a combination model of different media (text, audio, video, etc.) is exciting to contemplate.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect multiple query representations on information retrieval system performance. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346, New York, NY, USA, 1993. ACM Press.

[2] F. C. Gey and A. Chen. TREC-9 cross-language information retrieval (English-Chinese) overview. In *The Ninth Text REtrieval Conference (TREC-9)*, pages 15–24. National Institute of Standards and Technology, October 2001.

[3] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM Press.

[4] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval.

Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5):735–750, 2004.

[5] M. Montague. *Metasearch: Data Fusion for Document Retrieval*. PhD thesis, Dartmouth College, 2002.

[6] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information System*, 9(3):187–222, 1991.