Applying Machine Translation Resources for Cross-Language Information Access from Spoken Documents

Gareth J. F. Jones
Department of Computer Science, University of Exeter
Old Library, Prince of Wales Road, Exeter EX4 4PT, U.K.
email: G.J.F.Jones@exeter.ac.uk

Abstract

The current expansion in collections of natural language based digital documents in various media and languages is creating challenging problems for accessing the information contained in these documents. This paper introduces an approach to cross-language information access for spoken documents which combines machine translation, information retrieval and speech recognition. The paper reports an initial experimental investigation into the retrieval of English language spoken video-mail messages using French language search requests.

1 Introduction

The current rapid expansion in the availability of multilingual material, increasingly contained in different media, is creating many demands for technology to automate access to information contained within these archives. Solutions to these information access problems cannot rely on a single technology, but must integrate a number of existing technologies. For example, access to material in a multilingual collection of documents containing spoken, electronic text and handwritten documents, could require information retrieval (IR), machine translation (MT), speech recognition (SR), and optical character recognition (OCR). In order to facilitate efficient document browsing techniques from text summarisation and information visualisation might also be required. Research in all these individual areas has been ongoing for many years, and interest in all of them has increased significantly in recent years as information processing demands have expanded. This paper describes initial work on a project to explore the effectiveness of MT resources for cross-language spoken document retrieval and for browsing of retrieved documents. Experimental results indicate that, while retrieval effectiveness is impaired by errors arising in both language translation and SR, the overall system still achieves reasonable performance.

The remainder of this paper is organised as follows. Section 2 introduces the topic of Information Access in more detail, Section 3 and Section 4 respectively outline the SR and language translation methods used in this work. Section 5 describes the IR techniques used in this investigation, while Section 6 and Section 7 describe the experimental data-set and results. Finally, Section 8 gives concluding comments and outlines directions for further work.

2 Information Access

Figure 1 shows an example of a complete system for a Cross-Language Information Access (CLIA) system for spoken documents. The following sections outline the concepts and components of this figure, starting with a standard monolingual information retrieval system and expanding to the complete CLIA system.

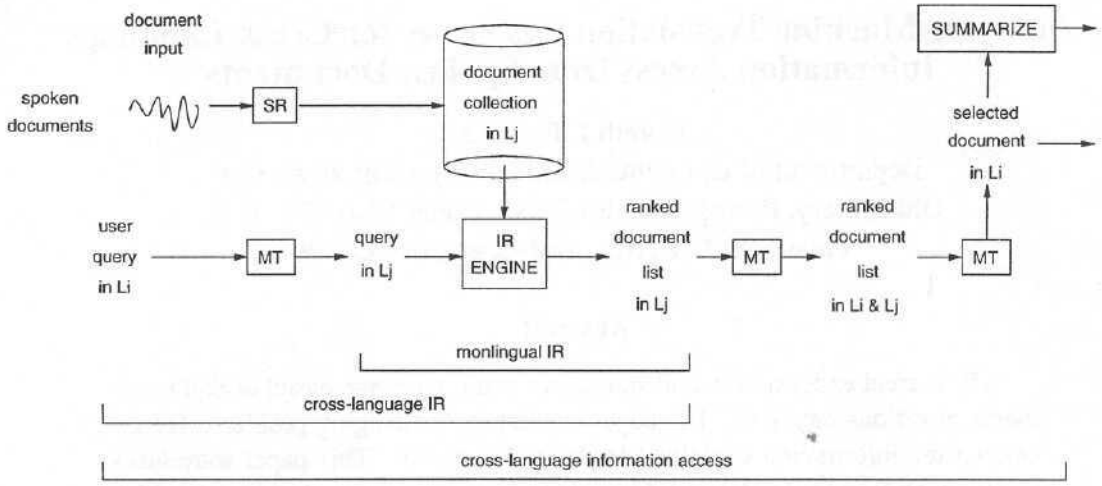

Figure 1: System diagram for a simple Cross-Language Information Access (CLIA) system for spoken documents.

2.1 Information Retrieval

When using an *information retrieval* (IR) system a user is primarily interested in *accessing* information contained in documents indexed by the retrieval system. IR itself is usually taken to be the location and retrieval of documents potentially relevant to a user's information need. As shown in Figure 1, in standard monolingual IR the user enters a search *request*, usually in natural language, in response to which the IR system returns a ranked list of potentially relevant documents in the same language as the request. It is assumed within this scenario that the user will be able to express their information need in the same language as the documents and that they will be able to decide if a retrieved document is relevant and, if it is, to extract useful information from the document. As shown in Figure 1 spoken documents must be indexed prior to retrieval by using a speech recognition (SR) process. SR typically makes errors in indexing which will ultimately impact on retrieval performance.

2.2 Cross-Language Information Retrieval

Monolingual IR can be extended to the retrieval of documents in a language j using search requests in a different language i, as illustrated in Figure 1. This process is referred to as Cross-Language Information Retrieval (CLIR). In order to perform CLIR either the requests or the documents must be translated. Whether to translate the requests or the documents has been the focus of a number of research studies. However, it is generally held that in practice request translation is the more viable option, and this approach is followed in this paper.

In a truly multilingual collection documents could originate in many different languages, and requests could be posed in many different languages. Issues relating to multilingual collections are not considered further here, and documents and requests are restricted to a single language pair of English and French respectively.

The role of translation in CLIR is essentially to bridge the gap between surface forms of terms in the request and document languages. The three major practical challenges in CLIR are: *coverage*: providing sufficient bilingual knowledge; *disambiguation*: how to identify conceptually different forms from the set of possible translations of a request word; and *synonym selection*: how to identify conceptually

equivalent forms of a translation. Machine Translation (MT) using deep linguistic analysis is a core-technology for providing solutions in all of these areas. The main limitations which arise in adapting MT to IR are in the coverage of the bilingual dictionaries and in the amount of context available in short IR search requests, where it is difficult for linguistic analysis to succeed. These problems are non-trivial and increase as the scope of language to be processed increases. While the focus of this paper is on the use of MT techniques in CLIR, it is important to place this in the overall context of CLIR research. This section gives a brief outline of translation techniques which have been applied to CLIR, highlighting the advantages and disadvantages of each. Request translation methods for CLIR fall into three categories: *dictionary term lookup* (Hull & Grefenstette 1996) (Ballesteros & Croft 1998); *machine translation* (Ballesteros & Croft 1998) (Carbonell, Yang, Frederking, Brown, Geng & Lee 1997) (Jones, Sakai, Collier, Kumano & Sumita 1999) and *parallel-corpora* methods (Sheridan & Ballerini 1996) (Carbonell et al. 1997).

Dictionary Term Lookup (DTL) DTL is the simplest translation method and involves simply replacing each request word by all its possible translations from a bilingual dictionary. The main disadvantage of DTL is the high degree of ambiguity often introduced into the translated request which can significantly degrade retrieval performance. Despite this problem, DTL is the most widely used method for CLIR because of its ease of implementation and the limited access to effective full MT resources for many language pairs (Hull & Grefenstette 1996) (Ballesteros & Croft 1998).

Machine Translation (MT) MT uses all available linguistic resources to calculate a single best possible translation of the whole request. This has the advantage that the MT system attempts to resolve all ambiguity and generate the single best translation of the concepts conveyed by the request. It has been argued in some of the literature that the shortness and lack of linguistic structure in typical search requests and domain dependence issues (Hull & Grefenstette 1996) mean that MT is unsuitable for CLIR and that dictionary methods should be favoured. However, recent research has challenged this hypothesis (Jones et al. 1999) (Ballesteros & Croft 1998) and indicates that MT can often outperform DTL in CLIR.

Parallel Corpora Methods (PCM) In PCMs two parallel (or usually more strictly comparable) document collections in the request and document languages are used to translate the query. These methods have shown some promise (Sheridan & Ballerini 1996) (Carbonell et al. 1997). However they rely on the availability of a suitable parallel document collection for the language pair, with the further complication that the documents must be related to the subject matter of the request. While not discounting the potential utility of parallel corpora methods, they are not considered further in this paper.

At a linguistic level the DTL and MT approaches can be regarded as extremes of translation complexity. The investigation reported in this paper compares CLIR performance using DTL and MT with monolingual baseline performance.

2.3 Cross-Language Information Access

In a CLIR system once the documents have been retrieved the user must still extract information from them to address their initial information need. However the user may often not have a sufficient level of fluency in the document language to do this. A CLIA system adds additional stages to the document retrieval process to assist users in accessing information. Thus, in the example shown in Figure 1 after retrieval, MT is applied again to translate the title and perhaps first phrase of each of the retrieved documents into language i. This augmented ranked list is then shown to the user. Individual documents chosen from this list are then passed to the MT system for full translation. Optionally these documents can then be passed to a Summariser to reduce the amount of time spent browsing the document contents. Overall the objective in design of an CLIA is to maximise its performance in terms of satisfying the user's information need. The only previously reported research on cross-language spoken document retrieval (CLSR) is (Sheridan, Wechsler & Schauble 1997) which uses a PCM translation technique. This paper describes a new CLSR investigation exploring the use of MT and DTL for CLSR.

3 Indexing Spoken Documents

In order to retrieve spoken documents, their contents must be indexed using a speech recognition system. The experiments described in this paper use standard speech recognition techniques based on Hidden Markov Models (HMMs), which are the basis for most current speech recognition systems. In speech recognition an HMM is a statistical representation of a speech event such as a word or subword phone. HMM parameters are typically trained on a large corpus of labelled acoustic speech data. The experiments in this paper index the spoken documents using a large vocabulary speech recognition (LVR) system which attempts to transcribe the complete spoken contents of the data. Note that since the recognition vocabulary is *large* rather than *open*, spoken words outside this vocabulary cannot be used for indexing and will by definition be indexed incorrectly. These words are referred to as Out-of-Vocabulary (OOV). The LVR system in this work uses the HTK tool set developed at Cambridge University (Young, Woodland & Byrne 1993). This is a powerful and flexible set of software tools for developing HMM applications. In these experiments, subword acoustic triphone models were trained on the WSJCAM0 British English speech corpus (Jones, Foote, Sparck Jones & Young 1996). Ideally a suitable language model would be built for the LVR system using a large archive of text material typical of the application domain. Unfortunately since there was no suitable archive of this type available, a standard WSJ 20K bigram language model from MIT Lincoln Labs was used (Jones et al. 1996). This model was constructed using a set of articles from the Wall Street Journal.

4 Translation Strategies

As stated earlier this investigation of CLIR focuses on the translation of search requests rather than documents. In the investigation requests were posed in French and then translated into English. Two translation strategies were explored: DTL and MT.

DTL was kindly performed by the Xerox Research Centre Europe, Grenoble using their bilingual dictionary developed for CLIR (Hull & Grefenstette 1996). For this translation each French search word is replaced by each possible translation into

English. MT was carried out using the *Globalink Power Translator Pro Version: 6.4* system. Power Translator Pro produces the best possible overall translation of the request. Examples of both of these translations are shown in the experimental section.

5 Information Retrieval Techniques

For the experiments in this paper, standard IR indexing and matching techniques were applied to the transcription files of the spoken documents. The experiments make use of the Okapi BM25 probabilistic *combined weight* (*cw*) (Robertson, Walker, Jones, Hancock-Beaulieu & Gatford 1995). The BM25 *cw* term weight is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i)$ is the standard collection weight (often referred to as the inverse document frequency (idf) weight) calculated as,

$$cfw(i) = \log \frac{N}{n(i)}$$

where N is the total number of documents and $n(i)$ is the total number of documents containing i, $tf(i,j)$ is the frequency of term i in document j, and $ndl(j)$ is the normalised length of document j, calculated as,

$$ndl(j) = \frac{dl(j)}{\text{Average } dl \text{ for all documents}},$$

where $dl(j)$ is the length of document j. $K1$ and b are empirically selected tuning constants for a particular collection. $K1$ is designed to modify the degree of effect of $tf(i,j)$, while constant b modifies the effect of document length. High values of b imply that documents are long because they are verbose, while low values imply that they are long because they are multitopic. The request-document matching score is computed by summing the weights of search items present in both the request and the document.

6 Experimental Data

The experiments reported in this paper use the Video Mail Retrieval (VMR) Collection. The VMR data set was developed as part of a project at Cambridge University (Jones et al. 1996). Fifteen speakers (11 men and 4 women) recorded 5 hours of spontaneous speech messages. Each speaker provided 20 spontaneous speech messages in response to 5 prompts chosen from 4 categories. The resulting 300 messages, along with their manual text transcriptions, serve as the test corpus for the retrieval experiments. The messages, though prompted, are fully spontaneous and contain a large number of disfluencies such as "um" and "ah," partially uttered words and false starts, laughter, sentence fragments, and informalities and slang ("'fraid" and "whizzo"). The VMR1 message set is very small by text retrieval standards, but is a useful starting point for experiments in CLIA for spoken documents since detailed monolingual results for this collection have been reported previously (Jones et al. 1996). Data was

recorded at a 16 kHz sampling rate using a Sennheiser HMD 414 head-mounted microphone, the standard for most research in speech recognition. For speech model training and recognition, the acoustic data was parameterized into a spectral representation at a 100 Hz frame rate.

The LVR system outlined in Section 3 using the WSJ triphone set and bigram language model taken together yielded a 53% word recognition accuracy rate. This is low compared to read speech, where accuracy rates often exceeds 90% in a limited domain, but is respectable given the difficulty of the spontaneous VMR task. Ongoing research in speech recognition is yielding continued improvements in recognition performance.

6.1 Retrieval Collection VMR1b

The VMR1 message collection forms the basis of the VMR1b retrieval collection. VMR1b combines the messages with a set of 50 requests and relevance assessments. VMR1b was obtained by asking users to generate natural language search requests in English, as stimulated by a prompt for each message category. A suitable relevance assessment subset was formed by combining the 30 messages in the category to which the original message prompt belonged, with the 5 messages from outside the category having the highest text retrieval scores. Users then marked messages from this subset which were relevant to their request. This gave 10.8 highly relevant documents on average per request. The requests average 12.0 words in length. After removing the standard van Rijsbergen stop words (van Rijsbergen 1979), an average of 7.4 content words remain. On average 6.6 of the words are found in the 20K LVR vocabulary. Even though for this collection there is only about 1 such word per request, OOV terms are likely to be domain specific and as shown in (Jones et al. 1996) are potentially important in maximising retrieval performance. Next all request words and hypothesised document contents are suffix stripped to search *terms* to encourage matching of different word forms using the standard Porter algorithm (Porter 1980). For example, given the following user search request,

```
Show all messages about privacy concerns due to always-
active video cameras.
```

the following query was obtained:

```
show messag privaci concern due alwai activ video camera
```

Word fragments such as "messag" are the suffix-stripped word forms.

6.2 VMR1b CLIR Collection

In order to form an experimental CLIR collection from the monolingual VMR1b collection three native French speakers translated the English search requests into French. The previous example was translated as follows,

```
Montrer tous les messages traitant des soucis d'intimité
en raison des caméras vidéo toujours actives.
```

For the CLIR experiments these French requests were then translated back into English using the DTL and MT methods outlined in Section 4. Using DTL this request was translated as follows,

```
point point out show reveal appear message get message
```

```
across. treat negotiate do or make deal deal with pro-
cess problem marigold intimacy cosiness depths private
life privacy reason ratio movie camera camera video al-
ways anyway still vigorous working person buoyant stimu-
late stir up stoke speed up active activate
```

This example illustrates the classic features of the DTL method for CLIR. The translated request typically contains most of the words found in the original English request along with many others which are not related to the subject of the request. It is often found that these additional terms can impact significantly on CLIR performance since they may have high $cw(i, j)$ weights. Using the MT system the request is translated as follows,

```
to show all the messages always treating the worries the
intimacies camcorders active.
```

which although it captures the essential meaning of the original English request uses rather different vocabulary. The weakness of this translation method for CLIR is thus that there may be search term mismatch between the query and the document.

7 Retrieval Experiments

This section presents experimental retrieval results in a series of comparisons. Results are shown using retrieval precision at ranked cutoff of 5, 10, 15 and 20 documents, and TREC average precision. TREC average precision is calculated as follows. For each query starting from the top of the ranked retrieval list each relevant document is located in the list and the precision value calculated at that rank position. The precision values for all relevant documents for the query are then averaged. Finally, the TREC average precision is calculated by taking the average of the individual query average precision values across the whole query set. The empirical cw parameters were selected as $K1 = 1.0$ and $b = 1.0$ based on previous work (Jones et al. 1996). With a small test collection such as VMR1b specific figures are neither reliable nor significant: the emphasis is therefore on the general picture that emerges from the results.

The first experiments give reference performance for monolingual text retrieval for all the query terms, versus only those terms in the 20K vocabulary, and then with the spoken documents using LVR for content indexing. The latter experiments show CLIR performance first with text retrieval again with open vocabulary, then with 20K vocabulary and finally with LVR. These experiments compare translation using DTL and MT.

7.1 Monolingual Experiments

Table 1 shows monolingual retrieval performance for the VMR1b collection. It can be seen that performance is degraded by the limitation to the 20K text vocabulary, and that there is a further reduction when indexing using LVR due to recognition errors.

7.2 Cross-Language Information Retrieval

Tables 2 and 3 show CLIR performance for DTL and MT request translation respectively. The lower rows show the difference in average precision with respect

		Open	20K Text	20K LVR
Precision	5 docs	0.388	0.333	0.308
	10 docs	0.325	0.281	0.248
	15 docs	0.311	0.274	0.239
	20 docs	0.296	0.264	0.218
Av Precision		0.371	0.327	0.265
% change		—	-11.9%	-28.6%

Table 1: Monolingual retrieval performance for the VMR1b collection.

		Open	20K Text	20K LVR
Precision	5 docs	0.258	0.246	0.200
	10 docs	0.233	0.210	0.190
	15 docs	0.226	0.199	0.168
	20 docs	0.213	0.196	0.155
Av Precision		0.253	0.232	0.177
% change clir		—	-8.3%	-30.3%
% change mono		-31.8%	-37.5%	-52.3%

Table 2: CLIR performance for the VMR1b collection using DTL request translation.

		Open	20K Text	20K LVR
Precision	5 docs	0.304	0.271	0.267
	10 docs	0.298	0.254	0.225
	15 docs	0.265	0.235	0.208
	20 docs	0.245	0.215	0.187
Av Precision		0.319	0.284	0.238
% change clir		—	-11.0%	-25.4%
% change mono		-14.0%	-23.5%	-35.8%

Table 3: CLIR performance for the VMR1b collection using MT request translation.

to open vocabulary text for CLIR and monolingual IR. It can be seen that averaged across the VMR1b query set that MT performs significantly better than DTL for all text and LVR indexing. As would be expected, in both cases performance is degraded for CLIR compared to monolingual IR.

8 Conclusions and Further Work

This paper has described an initial investigation into Cross-Language Information Retrieval (CLIR) for spoken documents. Experimental results on the VMR1b collection show that retrieval performance is affected by both the vocabulary limits and recognition errors of LVR. Also CLIR performance is degraded by problems in translation of search requests. The results here suggest that on average MT is a better method for translation than simple DTL methods.

There are several directions in which further work is currently directed. First, a much larger retrieval test collection for CLSR experimentation is being developed.

Second, the LVR system used for comparative purposes with existing results in this paper needs to be replaced with a more sophisticated system with a larger recognition vocabulary. Third, the effectiveness of feedback methods such as pseudo relevance feedback (Ballesteros & Croft 1998) needs to be explored. Finally, a formal investigation is required into the use of MT and Summarisation for CLIA both for textual and spoken documents.

One interesting topic which it will be difficult to explore formally for practical reasons, but is possibly important is the choice of vocabulary in the LVR and MT systems. Development of dictionaries is an important topic in both LVR and MT systems. The effectiveness of CLSR systems may be improved if these dictionaries are developed to be complementary, rather than independently.

Acknowledgements

I am grateful to the Foreign Language Centre, University of Exeter for the use of the Power Translator Pro software and to the Xerox Europe Research Laboratory, Grenoble for providing the bilingual dictionary request translations.

References

Ballesteros, L. & Croft, W. B. (1998), Resolving Ambiguity for Cross-Language Retrieval, *in* 'Proceedings of ACM SIGIR 98', ACM, Melbourne, pp. 64–71.

Carbonell, J., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y. & Lee, D. (1997), Translingual Information Retrieval: A Comparative Evaluation, *in* 'Proceedings of IJCAI 97', Nagoya, pp. 708–714.

Hull, D. A. & Grefenstette, G. (1996), Querying across languages: A dictionary-based approach to multilingual information retrieval, *in* 'Proceedings of ACM SIGIR 96', ACM, Zurich, pp. 49–57.

Jones, G. J. F., Foote, J. T., Sparck Jones, K. & Young, S. J. (1996), Retrieving spoken documents by combining multiple index sources, *in* 'Proceedings of ACM SIGIR 96', ACM, Zurich, pp. 30–38.

Jones, G. J. F., Sakai, T., Collier, N. H., Kumano, A. & Sumita, K. (1999), A comparison of query translation methods for english-japanese cross-language information retrieval, *in* 'Proceedings of ACM SIGIR 99', ACM, San Fransisco, pp. 269–270.

Porter, M. F. (1980), 'An algorithm for suffix stripping', *Program* **14**(3), 130–137.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M. & Gatford, M. (1995), Okapi at TREC-3, *in* D. K. Harman, ed., 'The Third Text REtrieval Conference (TREC-3)', pp. 109–126.

Sheridan, P. & Ballerini, J. P. (1996), Experiments in Multilingual Information Retrieval using the SPIDER system, *in* 'Proceedings of ACM SIGIR 96', ACM, Zurich, pp. 58–65.

Sheridan, P., Wechsler, M. & Schauble, P. (1997), Cross-Language Speech Retrieval: Establishing a Baseline Performance, *in* 'Proceedings of ACM SIGIR 97', ACM, Philadelphia.

van Rijsbergen, C. J. (1979), *Information Retrieval*, 2nd edn, Butterworths, London.

Young, S. J., Woodland, P. C. & Byrne, W. J. (1993), *HTK: Hidden Markov Model Toolkit V1.5*, Entropic Research Laboratories, Inc., 600 Pennsylvania Ave. SE, Suite 202, Washington, DC 20003 USA.