CLIR as Query Expansion as Logical Inference

Jian-Yun Nie
DIRO
University of Montreal
Email: nie@iro.umontreal.ca

Abstract
Cross-language IR (CLIR) has been usually described as two separate steps: query translation and query evaluation. The uncertainty of the first step is not always integrated in the calculation of the correspondence between a document and the original query. In this paper, we try to develop a unified framework in which both steps are integrated. It will turn out that CLIR is a special case of query expansion, and a query expansion process can usually be formulated as a logical inference. The unified framework is based on logical inference. The final formula corresponds well to our previous CLIR experiments that have been done without strong theoretical justification. This paper also provides some justification for them.

1. Introduction

There have been many experiments on Cross-Language Information Retrieval (CLIR) during the last years. In most cases, people concentrated on the translation of queries from a language to another, using one of the following methods: Machine Translation (MT) system, bilingual dictionary, or probabilistic translation model. All these methods have been described in an ad hoc way. CLIR is formulated as two separate tasks: query translation and monolingual IR. In this paper, we try to integrate the tasks in the same framework. Both problems will be formulated as a logical inference process. The whole CLIR process consists then in two steps: finding related query forms using inference, and direct matching of documents from each query form.

With some simplification assumptions, we can arrive at the following evaluation formula:

$$P(d \rightarrow q) = \sum_{q'} [P(d \rightarrow q') * P(q' \rightarrow q)]$$

in which d and q represent a document and a query, q' represents a possible alternative (or related) form of the query, and P is a (probability) function. This formula has a close relationship with the following probabilistic formula:

$$P(q|d) = \sum_{q'} [P(q'|d) * P(q|q')]$$

These formulas correspond to the current approaches used in CLIR and query expansion.

In the remaining of this paper, we will review the ideas of query expansion and CLIR, and arrive at the above two formulas. Our previous experiments on CLIR and query expansion will be used to support this formulation.

This article does not propose a completely new model for IR. Rather, we try to express some previous ideas in a simple, unified framework.

2. CLIR and query expansion practice

Let us first review briefly the current practice on CLIR and query expansion.

2.1. Cross-language IR

The core problem of CLIR (in addition to the common problems of IR in general) is to translate queries from a language to another. There have been mainly three approaches for this: using a machine translation (MT) system, using a bilingual dictionary, or using parallel texts.

MT systems seem to be a straightforward choice for query translation. For each query *q*, an MT system will give a unique translation *q'* for it. In some cases, the translation is reasonable. In other cases, the translation may depart from the original query. For example, the query "What effects has logging had on desertification?" is translated by Systran (Systran) to "Quels effets l'enregistrement a-t-il eus sur la desertification?" in French in which "logging" has been taken in the sense of "logging in a computer". Therefore, the generated new query *q'* is not an equivalent to the original *q*. However, there is no means for us to measure the uncertainty produced during the translation of *q* to *q'* by an MT system. The proposed approach has to use *q'* as an equivalent to *q*, and considers the relationship between a document *d* and *q'* as a close approximation of that between *d* and *q*. We can see here that the translation process is separated from the search process.

On the other hand, using a bilingual dictionary, the original query is translated word by word: Each word is considered to be independent from the others in the query. We are faced with the problem of multiple translation choices because each word usually has several possible translations. Several approaches have been used to cope with this problem: One can choose the first translation, assuming that it is the most common translation of the word; One can also use all the translation words together and associate to them an equal weight; Finally, the translation words can be associated with a weight according to their distribution within the document collection (e.g. the more a word appears in the collection, the higher its weight is).

Finally, by using a set of parallel texts (texts with their translations), we can extract translation relationships from them. The principle is that, the more two words (or phrases) co-occur in parallel texts (or sentences), the more they are translation each for another. This principle has been implemented in two ways. In (Yang 1998), the original query is first used to retrieve the texts in the source language from the parallel corpus that match the query. A set of keywords is then extracted from the corresponding texts in the target language. It is used as a query translation to retrieve documents in the target language from the document collection. More often, the parallel texts are used to train a statistical translation model (Nie et al. 1999). For CLIR purposes, the IBM model 1 is usually used (Brown et al. 1992). The core of the model is a probability function $P(t|s)$ that gives the probability to translate a source word *s* to a target word *t*.

2.2. Query expansion

Query expansion works as follows: Given an initial user query, some new related words are added and this forms a new query. The addition of the new words extends the original query so that it has a wider coverage than the original query. As a consequence, more relevant documents are expected to be retrieved, and the recall ratio be increased. The key problem is to identify the appropriate words to be added. Otherwise, the new query will depart from the original query (in meaning). So an important question is what words should be added. Another important question is how they should be integrated into the new query.

How are new words integrated into the query?

Let us first examine this problem with respect to the two most used models: Boolean model and vector pace model. In Boolean model, the added words are put into disjunction with the original

query words. For example, if t is a word in the original Boolean query and t_1 is a related term to it, then t_1 is put into disjunction with t in the new query. In some cases, the added term is assigned a weight equal to that of the original term t. Thus, t is replaced by $(t \vee t_1)$. In other cases, the added term is assigned a lesser importance. So t is replaced by $(t \vee t_1^{\alpha})$ where $\alpha \leq 1$. The factor α has been assigned different functions. In some systems, it plays the role of a multiplication factor of the similarity obtained with t_1. That is, if a document's similarity to t_1 is v, then its similarity to t_1^{α} is $(\alpha * v)$. In some other systems, it is an upper bound of the similarity, i.e. the similarity to t_1^{α} is $min(\alpha, v)$.

In vector space model a related word is added into the corresponding vector dimension of the query vector if it does not exist in the original vector. If it exists, its weight is increased by a certain factor. The effect of query expansion in vector space model is similar to that in Boolean model. However, the new word is not considered as an alternative (expressed as disjunction) of the original word, but as a supplement to it.

Which words are added? - The use of thesauri

Automatic query expansion usually relies on a thesaurus (or pseudo-thesaurus), which stores a set of relationships between words or terms. Among the thesauri used, there are classical thesauri that are established manually (Miller 1990), or pseudo-thesauri that are established automatically using co-occurrence information (Rijsbergen 1977). In using a manual thesaurus, only strong relationships (e.g. the *is_a* relationships) are used (Voorhees 1994). In some cases, indirectly linked terms (related through more than one link) are also used, but with lower weights (Rada et al. 1991; Salton and Buckley 1988). As to pseudo-thesauri, co-occurrences considered are restricted within some context, which may be: document, paragraph, sentence or even some syntactic structure (Grefenstette 1992).

2.3. CLIR as query expansion

Although the CLIR problem has been formulated in a different way from query expansion, we can see a close relationship between them. This relationship is reflected with regard to two aspects: the principle and the knowledge used.

Principle

The query translation q' may be seen as an expansion to the original query q. The only difference lies in whether we keep the original query or not. However, this difference is not significant. If we associate a language marker with each keyword, the new query q' can be simply added to the original query q instead of replacing it. For example, if each index is a couple $<w, l>$ where w is a keyword and l its language marker, then the new query can be simply a mixture of words in two different languages. Similarly, documents have also to be indexed together with a language marker. In this way, query translation becomes exactly a query expansion. Therefore, we can conclude that query translation is a special case of query expansion.

It is interesting to notice that this query expansion view with language marker is an extension to the current CLIR approaches. In the current approaches, it is assumed that a document collection contains documents of the same language. The translation direction (from a language to another) is controlled manually. In reality, especially in the Internet environment, documents of different languages are mixed up. By adding a language marker with indexes, these documents may be indexed together. The retrieval in different languages may be done in a single pass. The requirement is that the language of each document may be recognized automatically. This is no longer a problem as there are several automatic language identifiers (e.g. (Isabelle 1997)) that can determine the language of a text at a very high precision (over 95% if the text is at least than a line long).

Knowledge

With respect to the knowledge used during query expansion or query translation, there is also a close relationship. In fact, a bilingual dictionary may be seen as a thesaurus: from a word, the translation relation leads to several other words in the other language. The use of parallel texts may be seen as a special case of exploiting document collection for term relationships using co-occurrences. As to MT, there is no strict correspondence in query translation approach. However, it is not difficult to see it as a heuristic means to derive related words from the original query.

With respect to these two aspects, we can then conclude that query translation is a special case of query expansion.

3. A unified framework

Let us assume that a document *d* is represented as a set of terms, or equivalently as a conjunction of terms. Each term corresponds to an atom. A query is a Boolean expression of terms. The relevance of a document represented by *d* to a query represented by *q* is determined by the logical implication $d \rightarrow q$. A logical system is characterized by a set of logical sentences (or its closure). If we represent it by *K*, then the relevance of *d* to *q* with respect to this system is expressed as $K \vdash d \rightarrow q$. If we have $K \vdash d \rightarrow q$, the document is said to be relevant. If we cannot prove $K \vdash d \rightarrow q$, it is irrelevant.

The element *K* denotes *system knowledge* that makes inference possible. Notice that K in the classical Boolean model is empty. Therefore, $K \vdash d \rightarrow q$ is proved only if d contains all the keywords required by *q.*

For the evaluation of $K \vdash d \rightarrow q$, we can inspire from the transitivity of classical logic implication:

$$A \rightarrow B \wedge B \rightarrow C \ \vdash \ A \rightarrow C$$

The evaluation of $d \rightarrow q$ with the system knowledge *K* may be done as follows (we remove *K* from all the following expressions):

$$d \rightarrow q' \wedge q' \rightarrow q \ \vdash \ d \rightarrow q$$

It means: if there is a new query *q'* such that the new query implies the original query, and that the new query is satisfied (implied) by a document, then we can say that the original query is also satisfied by the document.

As *q'* may be any query expression, we can re-write the above deduction as follows:

$$\bigvee_{q'} (d \rightarrow q' \wedge q' \rightarrow q) \ \vdash \ d \rightarrow q$$

Interpreting this formula in a context that involves uncertainty, we can define a function *P* so that:

$$P(d \rightarrow q) = P(\bigvee_{q'} (d \rightarrow q' \wedge q' \rightarrow q)) \qquad (1)$$

We notice that the right side of the equation (1) includes two factors:

- $P(d \rightarrow q')$ measures the *degree of (direct) satisfaction* of a query *q'* to the document *d*.
- $P(q' \rightarrow q)$ measures the *degree of relatedness* of the query *q'* to the original query *q*.

Now the question is to discompose the right side of the equation. In order to evaluate the right side of the formula, we have to define evaluation methods for logical conjunction and disjunction. In fuzzy set theory, many evaluation methods have been proposed for them. A general form - known as *triangular norm* Δ was suggested by (Dubois and Prade 1984) for the evaluation of conjunction. A triangular norm Δ: [0,1] x [0,1] $\rightarrow$ [0,1] is a function that verifies the following conditions (where $x, x', y, y', z \in [0,1]$):

1. $\Delta(x, y) = \Delta(y, x)$;
2. $\Delta(x, \Delta(y, z)) = \Delta(\Delta(x, y), z)$
3. If $x \geq x'$, and $y \geq y'$, then $\Delta(x, y) \geq \Delta(x', y')$.

Correspondingly, a disjunction is evaluated by a triangular co-norm ∇ that is defined as follows:

$$\nabla(x, y) = 1 - \Delta(1-x, 1-y).$$

The function *min* is a triangular norm. Its co-norm is *max*. Multiplication of real numbers is another triangular norm. Its co-norm is $(x + y - x*y)$. These two sets of functions are among the most used functions for logical operators in fuzzy set theory.

Using a triangular norm Δ and its co-norm ∇, we can further develop equation (1) as follows:

$$P(d \rightarrow q) = \nabla_{q'} [\Delta(P(d \rightarrow q'), P(q' \rightarrow q))] \qquad (2)$$

Let us define the evaluation of $P(d \rightarrow A^{\beta})$ as $\Delta(P(d \rightarrow A), \beta)$, i.e. the satisfaction of a query of the form A^{β} is a combination (by Δ) of the evaluation of A and the factor β. The above formula becomes the following:

$$P(d \rightarrow q) = \nabla_{q'} [P(d \rightarrow q'^{\,P(q' \rightarrow q)})]$$

Using the properties of triangular norm, we can easily obtain:

$$P(d \rightarrow q) = P(d \rightarrow \bigvee_{q'} q'^{\,P(q' \rightarrow q)}) \qquad (3)$$

This inference process is exactly the query expansion in Boolean model, i.e. the original query q is expanded to $\bigvee_{q'} q'^{\,P(q' \rightarrow q)}$, i.e. a disjunction of all the alternative query forms, together with their relationship with the original query. In practice, we often expand a query at term level, i.e. each term is expanded with the related terms. However, this version of query expansion can be easily derived from the above expression due to the properties of Boolean expressions and the triangular norm.

The problem can also be examined from a probabilistic point of view. A probabilistic IR model aims to evaluate the probability $P(R|d, q)$. Let us denote a related query by q', and assume that they are mutually exclusive and complete. Then we can obtain:

$$P(R|d, q) = \Sigma_{q'} P(R, q'|d, q) = \Sigma_{q'} P(R|d, q, q') P(q'|d, q)$$

The assumption that q's are mutually exclusive and complete is quite strong. It is not the case in practice. However, this is the assumption implicitly made behind most current query expansion methods.

In addition, we assume that q' is a good approximation of q, so that

$$P(R|d, q, q') = P(R|d, q').$$

The derivation of q' only depends on q and is independent of d. Therefore:

$$P(q'|d, q) = P(q'|q)$$

We then have

$$P(R|d, q) = \Sigma_{q'} P(R|d, q') P(q'|q)$$

In this equation, $P(q'|q)$ denotes the relationship between the original query and a derived query, and $P(R|d, q')$ the relevance estimation of the document to the derived query.

This equation has some relationship with equation (2). From equation (2), if we use multiplication as triangular norm, and make some simplification, we can obtain:

$$P(d \rightarrow q) = \sum_{q'} [P(d \rightarrow q') * P(q' \rightarrow q)]$$

The simplification we made is in fact the omission of $x*y$ in the evaluation of disjunction ($x + y - x*y$).

Most query expansion approaches expand terms (or keywords) instead of the entire queries. If we take the same approach, we can have:

$$P(R|d, q) = \Sigma_{t'} P(R, t'|d, q) = \Sigma_{t'} P(R|d, q, t') P(t'|d, q)$$

where t' is an extended term. Again, we assume that all the t's are mutually exclusive and complete (this assumption may seem more acceptable because it is the usual practice in IR. However, the essence of this assumption is the same as the previous one). We can make similar simplifications as follows:

$$P(R|d, q, t') = P(R|d, t')$$
$$P(t'|d, q) = P(t'|q)$$

i.e. the relevance of a document to a particular term t' is independent from the entire query q; and the derivation of an expanded term solely depends on the query q. Therefore:

$$P(R|d, q) = \Sigma_{t'} P(R|d, t') P(t'|q) \qquad (4)$$

$P(R|d, t')$ and $P(t'|q)$ may be interpreted respectively as term weight within a document, and the term importance for a query. Applied to CLIR, this formula allows us to consider the translation uncertainty as a factor in the estimation of relevance.

Equation (4) corresponds exactly to the CLIR approach we used (Nie et al. 1999). In our experiments, a set of parallel texts is used as training corpus for a translation model. This translation model allows us to calculate the probability $P(t'|q)$ – given an original query q, the probability to find word t' in its translation. So a query will be "translated" by a set of keywords t', together with its probability. Furthermore, our experiments used a vector space model in which the similarity is determined by inner product. The similarity obtained is exactly the same as defined in equation (4).

In our experiments, we found that the approach is effective. We were able to achieve a CLIR performance close to that using one of the best MT systems. This is another fact that supports the CLIR or query expansion approach developed so far.

5. Conclusion

In this paper, we tried to show that CLIR is a special case of query expansion, and query expansion may be formulated as a logical inference. An evaluation method of the uncertainty logical inference is defined using a general form of triangular norm. With some simplification assumptions, we arrived at a form of query expansion that is usually used in IR practice.

These simplification assumptions, however, are not all reasonable. They are quite strong. Are they a source of problem in query expansion and CLIR that did not surface until now? Can we explain some failures in query expansion attempts from these simplifications? These are the questions we are examining now.

The claim of this paper is that logical inference is at the center of many IR approaches. In this paper, a simple form of logic inference is used instead of complex ones as in (Crestani et al. 1998). This choice is made for the simplicity of its implementation. We do not claim, however, that this simple form of logical inference can cover all forms of inference that may occur in IR. However, this is out of the scope of this paper.

References

P. F. Brown, S.A.D. Pietra, V. D. J. Pietra, and R. L. Mercer, The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19: 263-312, 1992

F. Crestani, M. Lalmas and C.J. van Rijsbergen (eds.). *Information Retrieval, Uncertainty and Logics*, Kluwer Academic Publishers, pp. 17-38, 1998.

D. Dubois and H. Prade (1984). Fuzzy logics and the generalized modus ponens revisited. *Cybernetics and Systems: An International Journal*, 15: 293-331.

G. Grefenstette (1992). Use of syntactic context to produce term association lists. *Research and Development on Information Retrieval - ACM-SIGIR*, 89-97.

P. Isabelle, G. Foster et P. Plamondon, The *SILC project*, http://www-rali.iro.umontreal.ca/ProjetSILC.en.html, 1997

G. Miller (ed.) (1990). *Wordnet: an on-line lexical database*, *International Journal of Lexicography.*

J.-Y. Nie, M. Simard, P. Isabelle and R. Durand (1999). Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web, *Research and Development on Information Retrieval - ACM-SIGIR*, Berkeley, pp. 74-81.

R. Rada, J. Barlow, J. Potharst, P. Zanstra, and D. Bijstra (1991). Document ranking using an enriched thesaurus. *Journal of Documentation*, 47: 240-253.

C. J. van Rijsbergen (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33: 106-119.

G. Salton and C. Buckley (1988). On the use of spreading activation methods in automatic information retrieval. *11th ACM-SIGIR Conference*. pp. 147-160.

Systran. http://babelfish.altavista.digital.com/

Y. Yang, J.G. Carbonell, R.D. Brown, R.E. Frederking, Translingual information retrieval: learning from bilingual corpora, *Artificial Intelligence*, 103: 323-345, 1998.

E. M. Voorhees (1994). Query expansion using lexical-semantic relations. *Research and Development on Information Retrieval - ACM-SIGIR*, Dublin, 61-70.