[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Evaluating the Adequacy of a Multilingual Transfer Dictionary for the Cross Language Information Retrieval

Gregory Grefenstette
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, FRANCE
[Gregory.Grefenstette@xrce.xerox.com]

Abstract

Information exists on the Web in a number of languages. This situation has given rise to new line of research called Cross-Language Information Retrieval (CLIR), treating the problem of finding a document written in one language via a query written in another language. One of the important resources needed for this problem is set of bilingual dictionaries for producing queries in new languages. The two most important aspects of these bilingual dictionaries for CLIR are the coverage that the dictionary provides for domain-independent corpora, and the adequacy of the translations provided for finding relevant documents in the second language. In this paper, we present a number of evaluations of these aspects for a bilingual dictionary, available through the ELRA. These evaluations are run against large corpora used in the TREC information retrieval trials.

Introduction

In addition to all of its other qualities, the WWW has created the largest distributed database in the world. One interesting new characteristic of this database, contrary to most previously studied databases, is that it is inherently multilingual. This multilinguality can either be regarded as a source of noise, introducing spurious foreign-language documents in the response to a query search, or as a untapped source of richness. The latter view motivates the recently created off-shoot of Machine Translation and Information Retrieval research called Cross Language Information Retrieval (Grefenstette, 1998). Cross-Language Information Retrieval (CLIR) deals with the problem of retrieving documents written in a target language given queries written in a different source language. One of the primary problem of CLIR is finding the proper target language query terms given a source language query. Other than using parallel corpora to find translation equivalents, the main technique used by researchers in CLIR is to use an online dictionary to propose translations for query terms. In the this paper, we present techniques for evaluating the adequacy of online multilingual dictionaries for this CLIR task, and present the evaluation of a multilingual dictionary available through ELRA for the CLIR task proposed in the 1997 Text Retrieval Conference (TREC-6).

Evaluation Techniques

CLIR has an easier task than machine translation, since machine translation must only choose one translation alternative and reconstruct a syntactically correct output in the target language. CLIR can get away with proposing all the translation alternatives, including the correct one, for a given source language term in order for classical information retrieval methods to work. On the other hand, information retrieval systems, contrary to machine translation, must both be domain independent and assure wide coverage. This wide coverage requirement prevents any domain modeling and limits any domain-specific translation restrictions. For this CLIR task, the principal qualities of a multilingual dictionary are, then, coverage and correctness.

Corpus Coverage

For coverage, we can objectively measure the number of words in a given corpus that are translatable through the dictionary, modulo inflectional or relational morphology. We can examine how well a given publicly available multilingual dictionary covers the basic vocabulary of the TREC-6 multilingual test collection. We use the Basic Multilingual Lexicon, available from the ELRA[1], as our dictionary. This dictionary contains 37,600 senses translated across five languages: English, French, Spanish, Italian, and German.

The ELRA dictionary uses a numerical interlingua in order to provide translation equivalents. Each word or multiword expression (MWE) in the dictionary is presented on a separate line. Each line is divided into a universal identifier (a number from 1 to 37655), a one-letter language code, the word or multiword expression, and a part-of-speech code. Synonyms have the same universal identifier. For example, the English terms *agreement*, *arrangement*, *engagement* are all associated with the universal identifier 261 which maps to *accord* in French. A polysemous word can appear under different universal identifiers. For example, *agreement* appears also with the universal identifiers 8449 (in French, *convention).* In order to

[1] http://www.icp.grenet.fr/ELRA/cata/text det.html #basmullex

find translation alternatives, a simple technique is to find the entries corresponding to the words to translate, collect the universal identifiers, and then use these universal identifiers to access the corresponding terms in the target language. We wish to see how well this dictionary covers the words found in a general interest collection, such as though generally found in information retrieval experiments. For our reference corpora, we use the English, French and Spanish newspaper corpora that have been used in the US-government-sponsored Text Retrieval Conferences (TREC) to test non-English information retrieval and cross-lingual information retrieval. The characteristics of these corpora are given below in Table 1

Corpus	Source	Years	Mbytes
English	AP	88-90	750
Spanish	AFP	94-96	350
French	SDA	89-93	240

Table 1: Reference Corpora Characteristics.

Each corpus was part-of-speech tagged and lemmatized. The lemmatized forms (except for a small list of stopwords which were discarded, as is custom in information retrieval) were divided into five gross classes: nouns, verbs, adjectives, adverbs, and others, according to the part-of-speech tags returned for the word. The words in each group were sorted and the frequency of each lemma in the reference corpus was calculated and stored. The characteristics of the noun groups are given in the next tables.

Corpus	Mb	Total Nouns	Unique Nouns	Unique Lowercase
English	750	29 M	300 K	55 K
Spanish	350	14 M	138K	43 K
French	240	10 M	177 K	52 K

Table 2: Number of unique nouns and unique nouns composed only of lower-case letters, in each of the reference corpora.

In order to test the coverage of the ELRA dictionary for the reference corpora, for each lowercase noun we searched whether the noun was found in the ELRA multilingual dictionary, as an entry or as part of a multiword expression that was an entry. We decided to concentrate on lower case nouns in order to eliminate proper names from consideration[2]. Also, for expediency, we eliminated any noun containing a number or punctuation mark such as a period or a dash. Of the remaining common nouns (the last column in Table 2), their appearance in the bilingual dictionaries are described in the following tables. These tables labeled *English, French, and Spanish lower case nouns*, treat the nouns in the corpus by frequency. The most common noun in the English corpus is *year*, appearing 267,241 times. The data appearing in the second row of *English lower-case nouns* could be read as the following: "The first 2000 most frequent nouns in the English corpus account for 86% of all the occurrences
of nouns in the corpus. Of these 2000 unique nouns, 1813 (86%) are found as single entries in the ELRA dictionary. Another 64 (pushing up the coverage to about 91% of the 2000) are found only as part of a multiword entry on the ELRA dictionary; and another 15 (pushing the coverage to about 95%) words can be found as derivational equivalents of an ELRA entry."

English lower-case nouns

Nouns in Frequency Ranking	Cum % Noun Corpus	Noun Entry Present	Present if MWE Included	Present Deriv Form
1000	74%	94%	97%	97%
2000	86%	91%	94%	95%
3000	92%	87%	91%	92%
4000	94%	84%	87%	89%
5000	96%	81%	85%	87%
10000	98.7%	67%	70%	74%
20000	99.7%	45%	47%	51%
50000	99.9%	20%	21%	23%

French lower-case nouns

Nouns in Frequency Ranking	Cum % Noun Corpus	Noun Entry Present	Present if MWE Included	Present Deriv Form
1000	77%	94%	96%	97%
2000	88%	89%	92%	94%
3000	92%	84%	87%	91%
4000	95%	81%	84%	88%
5000	96%	77%	80%	84%
10000	98.6%	61%	64%	69%
20000	99.5%	40%	42%	47%
50000	99.9%	18%	19%	22%

[2] In reality, coverage of proper names in a bilingual dictionary is a task in itself since transliterated proper names are written differently in French, English and Spanish, e.g.. Yeltsine, Elstine, Yelstin. We did not do such an evaluation.

Spanish lower-case nouns				
Nouns Freq Ranking	Cum % Noun Corpus	Noun Entry Present	Present if MWE Included	Present Deriv Form
1000	78%	93%	95%	NA
2000	89%	87%	90%	NA
3000	94%	83%	86%	NA
4000	96%	78%	82%	NA
5000	97%	74%	77%	NA
10000	99.2%	57%	61%	NA
20000	99.7%	35%	38%	NA
40000	99.9%	19%	20%	NA

Tables 3,4,5: ELRA dictionary coverage of lemmatised lower-case nouns appearing in the corpora. Results are presented in function of the frequency of appearance of nouns in the corpus. *Cum% Noun Corpus* means what percent of all nouns in the corpus are covered by the N most frequent nouns in the corpus. *Noun Entry Present* means that this is a one-word ELRA dictionary entry for these noun. *Present if MWE included* means that there exist at least a multiword expression (MWE) entry including this noun. *Present Deriv Form* means that at least a derivational variant of the noun appears as a dictionary headword.

The chance of pulling a random noun out of the English corpus is 92%. This is because when one pulls a random noun out of the corpus, one is likely to pick a frequently occurring noun, and these nouns are well covered by the dictionaries: for all three languages the 1000 most frequent nouns are all covered by more than 93%.

Examples of common English words not found in the dictionary (with their frequency in the corpus) are chairman (34587), aide (10698), bushel (9873), lawmaker (9659), investigator (9456), pentagon[3] (8953), ounce (8111), ... Examples of common English words which appear in the ELRA dictionary but only as parts of multiword expressions are: the lemma *fund* which appears in the corpus 22056 times but which only appears in the ELRA dictionary as parts of the multi-word expressions (MWE) *reserve funds,* and *public funds;* the lemma site appearing only in *construction sites, nest sites, archeological sites;* the lemma *aircraft* which does not appear alone in the dictionary but as part of entries for *passenger aircraft, commercial aircraft, transport aircraft, aircraft propeller,* and *aircraft carrier.* Examples of corpus nouns which only appear in different derivational forms in the dictionary are *grocery* which appears in the dictionary only under the form of *grocer, peasant* appearing under the form *peasantry; statute* appearing under the form *statutory.* French examples of missing single word entries are: *taux, assemblée, jourrnée, bilan, scrutin, affrontement, homologue, provenance, votation, vaudaois.* Some of these words exist in multiword entries and some can be recaptured by using derivational variants.

Spanish examples of missing single word entries are *jornada, cargi, víspera, tonelada, asamblea, disparo, carretera, gestión, salario, acceso, penal, madrugada...* Again many of these words appear inside multiword entries (e.g., an entry for *de madrugada,* or for *reducción de jornada)* or as a derivational variant.

[3] The part-of-speech tagger lemmatises *Pentagon* to lower-case *pentagon.*

Translation Adequacy

The first problem in evaluating a bilingual dictionary for Cross Language information Retrieval (CLIR), examined in the previous section, is predicting how well a given multilingual dictionary will be able to cover the corpus by proposing at least one translation for each word which may appear a query. This gives a best-case analysis for CLIR, supposing that if translations exist for a word, then at least one of those translations is useful for information retrieval. A finer evaluation is provided by examining the translation coverage of the dictionary, that is, predicting how well the translations provided by the dictionary correspond to what is needed for information retrieval. In order to evaluate this, we will use the 25 manually translated TREC-6 cross language queries. For example, one of the queries that TREC has provided is the following:

Title: Les voitures solaires
Description: Des informations sur les voitures solaires.
Narrative: Un document pertinent contiendra des renseignements sur les recherches et le development des voitures solaires. Les voitures solaires font partie d'un effort pour freiner l'exploitation de carburants non renouvelables.

along with its English translation:

Title: Solar Powered Cars
Description: Information on solar powered cars.
Narrative: A relevant document will contain information on research and development on solar automobiles. Solar powered automobiles are part of an effort to popularize alternative energy sources to replace the continued exploitation of the world's finite fossil fuels.

If we consider that the French version of the query is the original query and that the English version is what we want the cross language information retrieval system to approach, then the translation process through the multilingual dictionary should produce at least the following English descriptors: *solar, power, car, research, development, automobile, popularize, alternative, energy, source, replace, continue, exploitation, world, finite, fossil, fuels.* The

evaluation is complicated a bit by the stemming process that is used in the information retrieval system. In our case, using the multilingual translation lexicon provides us with: *solar, car, research, development, automobile, fuel;* but not *fossil, popularize, power, world or finite.*

English to French CLIR terms found bytranslating lemmas

Query	Nouns Found	Adjs Found	Verbs Found	
CL1	12/18	4/6	0/2	62%
CL2	5/9	0/3	1/2	43%
CL3	4/9	4/6	1/3	50%
CL4	5/9	3/5	2/3	59%
CL5	9/12	5/7	1/2	71%
CL6	7/15	2/5	2/4	46%
CL7	8/15	0/1	4/4	60%
CL8	10/15	1/1	0/1	65%
CL9	,7/16	3/6	1/3	44%
CL10	8/10	2/3	0/2	67%
CL11	6/9	2/4	0/0	62%
CL12	5/10	2/4	0/1	47%
CL13	8/13	1/3	0/2	50%
CL14	4/6	2/3	1/4	54%
CL15	6/11	2/3	0/3	47%
CL16	9/14	2/3	1/2	63%
CL17	9/16	1/4	0/1	48%
CL18	7/8	2/5	0/3	56%
CL19	7/10	0/2	1/1	62%
CL20	9/12	2/6	2/4	59%
CL21	9/13	0/1	4/5	68%
CL22	6/8	1/3	0/1	58%
CL23	7/12	0/1	0/0	54%
CL24	8/16	0/2	2/3	48%
CL25	12/20	2/4	3/3	63%
total	61%	47%	44%	56%
total found using derivational stemming				60%

Table 6: ELRA dictionary coverage twenty-five TREC Cross Language Information Retrieval track queries. The queries are number CL to CL25. Each query was translated from an English version and the table entries show how many of the original french nouns, adjectives and verbs were found in the translated version..

The preceding table 6 shows the coverage provided by the ELRA dictionary for the queries (numbered CL1-CL25) in the most recent TREC Cross-Language Information Track, with query terms divided into gross classes of nouns, verbs and adjectives (most of the adverbs in the queries were also stopwords). The TREC Track provided both English and French versions of the queries. Starting from the English version, we lemmatized the English query. Each lemma was searched for in the ELRA dictionary, and all the translations for that lemma were collated to form a pseudo-French query. Then we checked whether the lemmatized French nouns, verbs, and adjectives from the original French query were found in the newly created pseudo-queries. The original French query CL25, for example, contained 20 nouns, 12 of which were found in the translations of query terms in the English version of the query terms through the ELRA dictionary; it contained 4 adjectives, 2 of which were found through the dictionary; and three verbs which were all found in the ELRA translations. For this query CL25, then, 63% of the original French query terms were found by the translation of English terms through the ELRA dictionary. Over all 25 queries, 53% of all terms in the original French queries were found again in the English-to-French translations. When we allowed derivational conflation of the terms before and after translation, the number of query terms found increased to 60%.

This is, again, only a best-case analysis of coverage, i.e. given the terms found in French query, we can find 60% of those same terms starting from a English version of the query, using the commercially delivered version of the ELRA multilingual dictionary. The translation method proposed does not eliminate any extra translations thrown in during the process. It is not clear, ahead of time (Hull & Grefenstette, 1996), whether these extras words will hurt (because importing noise) or help (because importing synonyms) the retrieval process.

Conclusion

In this paper we will have examined the adequacy of a large commercially available general-purpose multilingual translation dictionary for the task of cross language information retrieval. We calculated the coverage that the dictionary provides for three large domain-independent corpora in English, French, and Spanish. The dictionaries cover most of the frequently appearing words. We then calculated the coverage the dictionary provides for a standard test set of parallel queries, given the task of recreating French queries, using only the English versions of the queries and the given multilingual dictionary. For this last task, it is shown that in almost all of the 25 queries considered, at least half of the target French terms are provided by the dictionary.

References

Grefenstette, G. Hull, D. A., Gaussier, E. &. Schulze, B. M. (1997) "Xerox TREC-6 Site Report: Cross Language Text Retrieval" *TREC-6 Conference Working Notes*, INSTN, Gaithersburg, MD.

G. Grefenstette, ed. (1998) *Cross Language Information Retrieval* Boston, MA: Kluwer Academic Publishers:

Hull, D. & Grefenstette, G.(1996). "Querying across languages: A dictionary-based approach to multilingual information retrieval" In *SIGIR Proceedings*, Zurich, Switzerland, ETH.