Corpus-Based Query Translation for Cross-Lingual Information Retrieval

Ralf D. Brown
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15232-3890
USA

ralf+@cs.cmu.edu

Overview

1. Translation Methods:
   - Dictionary-based Term Translation
   - Example-based Term Translation
   - EBMT
2. Automated Dictionary Extraction
3. Evaluation
4. Results

Dictionary-Based Term Translation

For each word in the query:

1. look up the word in a general-purpose bilingual dictionary
2. add all translations of the input word to the translated query

Example-Based Term Translation

For each word in the query:

1. look up the word in a corpus-derived bilingual dictionary
2. for each translation of the input word,
    - compute its relative frequency from the frequency information in the dictionary
    - add the translation to the translated query 20*freq times (but at least once)

Example-Based Machine Translation

For each *line* in the query:

1. find all phrases which match against the source-language half of the example base
2. for each matching phrase:
   - attempt to find the translation by performing subsentential alignment
   - if the alignment is successful, add the translation of the phrase to the translated query

Automated Dictionary Extraction

The extraction process is a simple count-and-filter algorithm, consisting of the following steps:

- creating a table of bilingual co-occurrence counts
- filtering the table using a set of threshold tests

Two additional refinements are available:

- a positional bias for languages with similar word orders
- a minor modification for handling the highest-frequency words

The Co-Occurrence Table

The co-occurrence table is simply a two-dimensional matrix indexed by source-language words in one dimension and target-language words in the other. For each sentence pair in the corpus, increment the counts in the cells for the cross-product of the words in the two sentences. Thus, the sentence pair

```
the report on the
el informe sobre el
```

would cause the following elements in the co-occurrence table to be incremented:

```
c[the,el]       c[report,el]       c[on,el]
c[the,informe]  c[report,informe]  c[on,informe]
c[the,sobre]    c[report,sobre]    c[on,sobre]
```

Filtering

After constructing the co-occurrence table, all entries which do not pass at least one of the threshold functions are zeroed, and any remaining non-zero entries are output as probable translations.

The symmetric threshold is passed whenever

$$C[S,T] \geq threshold[C] \times count[S]$$

and

$$C[S,T] \geq threshold[C] \times count[T],$$

where $C[S,T]$ is the number of times source-language word S co-occurs with target-language word T and $threshold[C]$ is the threshold value selected by that co-occurrence count.

The asymmetric threshold is passed whenever

$$C[S,T] \geq thresh1[C] \times count[S]$$
and
$$C[S,T] \geq thresh2[C] \times count[T]$$

or

$$C[S,T] \geq thresh1[C] \times count[T]$$
and
$$C[S,T] \geq thresh2[C] \times count[S],$$

where $thresh1[C]$ and $thresh2[C]$ are the two separate limits of the asymmetric threshold.

Variable Thresholds

As shown on the previous slide, the threshold values are a function of the total co-occurence count. This permits adjustments for very low-frequency and very high-frequency words:

- very low-frequency words may have a fairly high co-occurrence ratio purely by chance (e.g a word which appears only twice in the entire corpus will have a co-occurrence of at least 0.50 with **any** word with which it co-occurs.
- very high-frequency words may have a high co-occurrence ratio with other very high-frequency words.

Sample Threshold Function:

C	Threshold
1	1.000
2	0.667
3	0.667
4	0.500
5	0.400
6	0.333
7	0.286
8	0.270
...	
1000	0.300
...	
2000	0.400
...	
3000	0.500

High-Frequency Words

The highest-frequency words occur so often that they correlate with each other sufficiently to pass the filtering tests. Thus, a slightly modified algorithm is needed.

To extract definitions for the highest-frequency words, a second pass is used, with the following differences:

- All but the highest-frequency words are omitted from the correspondence table.
- All sentences in the corpus which contain more than three high-frequency words are ignored.

When run for the 16 words occurring in 20% or more of all sentences, this modified algorithm extracts definitions for seven of them with no incorrect definitions.

Induced Definitions

(ASTUDILLO (ASTUDILLO 1))
(COMPLICACIÓN (COMPLICATION 9))
(REFUTA (REFUTES 3))
(OCTUBRE (OCTOBER 4350)(NOVEMBER 419)(SEPTEMBER 314)(DATED 314))
(ACOPIAR (OLS 4))
(CENTRALIZANDO (CENTRALIZING 4))
(K1N (OTTAWA 6)(ONTARIO 4)(K1N 2))
(CARTÓGRAFOS (CARTOGRAPHERS 2))
(CIEGOS (BLIND 18))
(CMINU (JUNIC 45))
(PÚAS (BARBED 8)(WIRE 8))
(DESMOTADO (COTTON 17)(GINNED 8))
(NUDO (IDICT 3)(NODE 3)(GORDIAN 3)(KNOT 3))
(TIEMPO (TIME 4723)(SAME 1357)(WHILE 1019))
(TRANSFORMARÁN (APPEASEMENT 2))
(PRISMA (PRISM 2))
(RESURGIMIENTO (RESURGENCE 36)(FASCISM 6))
(COEXISTIR (COEXIST 3))
(MAÍZ (MAIZE 31)(RICE 15)(METRIC 14)(CORN 10)(SORGHUM 9)(WHEAT 8)
(MILLET 6)(INSECTS 5))
(FECALES (FC 6)(COLIFORMS 3)(ML 3)(COLIFORM 3)(FECAL 3))
(ERRÓNEAMENTE (MISINTERPRETED 7)(WRONGLY 5))
(HORTALIZAS (VEGETABLES 31)(VEGETABLE 18)(FRUITS 15)(FRUIT 9))
(NOVIEMBRE (NOVEMBER 4810)(INTRODUCED 511)(OCTOBER 375))
(GANTCHEV (GANTCHEV 5))
(MASDIT (MASDIT 9))
(AGHRYMET (AGHRYMET 4))
(SABE (KNOWN 95)(KNEW 25))
(ABSOLUTAMENTE (ABSOLUTELY 65))
(O.O. (SPÓLKA 4)(AKCYJNA 2)(OGRANICZA 2)(ODPOWIEDZIALNOSCIA 2)
(SP.Z 2)(O.O. 2))
(ITINERARIOS (ITINERARIES 5)(PREDETERMINED 5))
(LAMENTAMOS (REGRET 34))
(DESPLAZADOS (DISPLACED 143)(EXTERNALLY 13)(INTERNALLY 12)
(AMERICANS 9))
(BARROCO (BAROQUE 3)(VIDEODISC 2))
(PRESUME (PRESUMED 7))
(DECLINAN (CAPTIVES 3))
(SHAHI (SHAHI 3)(MÜLLERSON 2))
(TIENDEN (TEND 96)(TENDED 24))
(SECTORIALISMO (SECTORALISM 2))
(MILOSEVIC (MILOSEVIC 11)(TUDJMAN 4)(KADIJEVIC 4))
(CINEMATECAS (PRINTS 2))
(DECIMALES (DECIMALS 2))
(|21B| (|21B| 2))
(BATOR (BATOR 7)(ULAN 4))

Size-Accuracy Tradeoffs

By setting the thresholds used in filtering to different values, a tradeoff between yield and accuracy may be tuned:

- Raising the thresholds reduces the number of incorrect/spurious translations generated, but reduces the size of the dictionary.
- Lowering the thresholds yields more definitions, but also increases the error rate.

The size-accuracy tradeoff needs to be tuned for each application. For EBMT subsentential alignments, a very low threshold (large dictionary, but also high error rate) proved best; for CLIR, a higher threshold (smaller dictionary, but considerably lower error rate) proved best.

Advantages of a Corpus-Derived Dictionary

- tuned to the way the corpus translates terms
- provides frequency information which is not available in general-purpose dictionaries
- provides some query expansion because monolingually-collocated terms are also included as translations
- can be tuned for best trade-off between accuracy and coverage

Evaluation Setup

Corpus:

- UN documents referring repeatedly to UNICEF (at least 10 times)
- ‘‘UNICEF’’ documents were subdivided along internal section boundaries to increase size of document collection
- human relevance judgements collected for each query over the collection

Metrics:

- Standard 11-point average precision
- Degradation from MLIR to CLIR

Baseline:

- SMART without relevance feedback (SMART.basic), optimized on the UNICEF collection

Sample Query Translations

Query:

water purification sanitation water supply project clean water personal hygiene health sanitation

Translated:

abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento agua agua agua agua agua agua agua agua agua agua agua agua agua abastecimiento agua agua agua agua agua agua agua agua agua agua agua agua agua agua abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento purificación abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento agua agua agua agua agua agua agua agua agua agua agua agua agua agua abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento abastecimiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento saneamiento salud higiene personales

Results

Method	Weighting	MIR	TIR	TIR/MIR
DICT	ntc.ntc	.4721	.2898	61%
DICT	ltc.ltc	.4306	.2340	54%
DICT	atc.atc	.3492	.1749	50%
EB-Term	ntc.ntc	.4721	.4318	91%
EB-Term	ltc.ltc	.4306	.3723	86%
EB-Term	atc.atc	.3492	.2803	80%

Comparison of EBT Dictionaries

Dictionary Threshold	TIR avgp (ntc.ntc)
0.10	0.3886
0.15	0.3936
0.20	0.4097
0.22	0.4235
0.25	0.4111
0.27	0.4318
0.30	0.3988

PanEBMT Corpus

Spanish-to-English corpus:

- ~685,000 sentence pairs -- ~265 megabytes
- UN Multilingual Corpus supplies most of the pairs
- 10,250 sentence pairs from Pan American Health Organization
- 552 sentence pairs from prior ARPA evaluations (newswire text)

Example:

```
Las fuentes de esos comentarios y
  recomendaciones son las suguientes:
The sources of these comments and
  recommendations are:
```

PanEBMT Architecture