Using surface-syntactic parser and Derivation from Randomness X-IOTA IR system used for CLIPS Mono & Bilingual Experiments for CLEF 2004

Gilles Sérasset
Laboratoire CLIPS-IMAG*
Grenoble France
Gilles.Serasset@imag.fr

Jean-Pierre Chevallet
IPAL-CNRS, I2R A*STAR
National University of Singapore
viscjp@i2r.a-star.edu.sg

Abstract

This document present work we have done for the CLEF 2004 participation. We promote the use of surface-syntactic parsing to extract indexing terms. We also promote the Derivation From Randomness weighting. For the bilingual part, we have tested reinforcement query weighting using an association thesaurus.

1 Introduction

In our previous participation at CLEF in 2003 [3], we have tested the use of an association thesaurus to enhance query translation. We have only use the association thesaurus to add some new terms to the proposed translation terms. In our current participation, we have tried another use of such a thesaurus: we do not enlarge the query, but rather we use it to modify the weighing of a given translated query. Our basic idea is the selection of the best term translation using query context and association thesaurus from the corpus. Last year, we have neglected the study of the matching function and the influence of the weighting scheme. For this participation we will also focus on this aspect: we have test the Derivation From Randomness (DFR) against Okapi measure and some other classical IR weighting. We also promote the use of a surface-syntactic parser. All documents are first transform by the parser. The stemming is then proposed by the parser. Finnish is an agglutinative language, and using such an NL parsing enable to correctly split the glued words into separated correct indexing terms.

The paper present first the training experiments performed on 2003 collection in part 2. In part 3 we discuss the monolingual results. Then, in part 4, we present the technique used for bilingual results and present hypothesis based on the results.

2 Training on monolingual run

In this part, we present some training we have achieved using monolingual corpus of CLEF 2003. We have mainly used the Finnish and French corpus. The purpose of this training is to select the best weighting scheme for the given CLEF document collection.

*This work is part of the PRISM-IMAG project devoted to high level indexing representation using interlingual graph formalism.

2.1 The underlying IR model

All experiments are grounded on the classic vector space model. Goal of experiment is to compare the probabilistic model of Okapi with Derivation From Randomness model, versus more classical weightings. This comparison will be done on two different languages.

Basically, the final matching process is achieved by a product between query vector and document matrix, which computes the Relevant Status Value (RSV) of all document against the query. For a query vector $Q = (q_i)$ with a dimension of t term $i \in [1..t]$, and a index document matrix of n documents $D_j = (d_{ij})$, $j \in [1..n]$, the RSV is computed by:

$$RSV(Q, D_j) = \sum_{i \in [1..t]} q_i * d_{ij}$$

We keep this matching process for all tests, the changes are in the documents and query processing to select indexing terms, and in the weighting scheme. We recall here the scheme that is inspired by the SMART system. We suppose the previous processing steps have produced a matrix $D = (d_{i,j})$. Usually, the value $d_{i,j}$ is only the result of term t_i counting in the document D_j, called term frequency tf_{ij}. Each weighting scheme can be decomposed in three steps: a local, a global and a normalization step. The local is related to only one vector. All these transformations are listed in table 1. For all measure we use the following symbols:

f_i	total number of term i in the corpus
d^*_{ij}	is a normalization of d_{ij}
λ_i	is the fraction $\frac{f_i}{T}$
T	is the corpus size : $T = \sum_i f_i$
df_i	the document frequency of i
d_{ij}	current value in the matrix
c	a constant for DFR
$L(D_j)$	the length of D_j equals to $\sum_t d_{ij}$
$awr(L(D_j))$	mean document length equals to $\frac{\sum_j L(D_j)}{n}$
n	number of document in the corpus
t	number of unique terms in the corpus
q_i	weight of term i of query q

n	$w_{ij} = d_{ij}$	none, no change
b	$w_{ij} = 1$	binary
a	$w_{ij} = \frac{0.5+0.5*d_{ij}}{max_i(d_{ij})}$	local max
l	$w_{ij} = ln(d_{ij} + 1)$	natural log
d	$w_{ij} = ln(ln(d_{ij} + 1) + 1)$	double natural log

Table 1: Local weighting

The global weighting is related to the matrix, and then it is a weighing which takes into account the relative importance of a term regarding the whole document collection. The most famous is the Inverse Document Frequency : Idf. The table 2 lists the global weighting we have tested. Okapi and DFR are not global weighting per se but rather complete weighting scheme themselves. In our X-IOTA system, they are computed at the same time than global weighting, and it is technically feasible to use them with a local and a final normalization. DFR is presented in the next part.

The Okapi measure described in [5, 4], uses the length of the document, the function $L()$, and also a normalization by the average length of all documents in the corpus, the function $A()$. This length is related to the number of indexing terms in a document. The Okapi measure uses 2 constants values called k_1 and b. Finally, the last treatment is the normalization of the final vector.

n	$w_{ij} = d_{ij}$	none, no global change
t	$w_{ij} = d_{ij} * log\frac{n}{df_i}$	Idf
p	$w_{ij} = d_{ij} * log\frac{n-df_i}{df_i}$	Idf variant for Okapi
O	$w_{ij} = \frac{(k_1+1)*d_{ij}}{k_1*[(1-b)+b*\frac{L(d_j)}{A(d_j)}]+d_{ij}}$	Okapi
R	(see below)	DFR

Table 2: Global weighting

n	$w_{ij} = d_{ij}$	none, no normalization
c	$w_{ij} = \frac{d_{ij}}{\sqrt{\sum_i d_{ij}^2}}$	cosine

Table 3: Final normalization

A weighting scheme is composed by the combination of the local, global and final weighting. We represent a weighting scheme by 3 letters. For example, `nnn` is only the raw term frequency. The scheme `bnn` for both documents and queries leads to a sort of Boolean model where every term in the query is considered connected by a conjunction. In that case the RVS counts the terms intersection between documents and queries. The `c` normalization applied to both document and query vector leads to the computation of the cosine between these two vectors. This is the classical vector space model if we use the `ltc` scheme for document and queries. The scheme `nOn` for the documents, and `npn` with the queries, is the Okapi model, and the use of `nRn` for document and `nnn` for the queries is the DFR model. For these two models, constants have to be defined.

Notice that the `c` normalization of the queries, leads to divide the RSV for this query by $\sqrt{\sum_i q_i^2}$. For each query this is a constant value which does not influence the relative order of answered document list. It follows that this normalization is useless for queries and should not be used. In the next section we briefly present the derivation from Randomness weighting that seems to give best results, and that we have used for all CLEF 2004 runs.

2.2 derivation from randomness (DFR)

This weighting scheme has been proposed by Gianni Amati in [1] (with a small error in the definition of the value $f^*_{t,d}$). Theoretical discussions about this approach can be found in [2]. Figure 4 sum up the results we obtain using this weighting scheme, on the CLEF2003 queries for training using the formula described in [1]. The formula is given by:

$$w_{ij} = (\log_2(1 + \lambda_i) + d^*_{ij} * \log_2 \frac{1 + \lambda_i}{\lambda_i}) * \frac{f_i + 1}{n_i * (f^*_{ij} + 1)} \quad (1)$$

The value d^*_{ij} is a normalization by the length $L(D_j)$ of the document D_j regarding the average size of all document in the corpus : $awr(D_j)$. A constant value c adjusts the effect of the document length in the weight.

$$d^*_{i,j} = d_{ij} * \log_2(1 + c * \frac{awr(L(D_j))}{L(D_j)}) \quad (2)$$

For this participation of CLEF, we have test this weighting scheme against another set of other computation. We present these results on Finnish and French collection.

2.3 Finnish IR

In these experiments, we have first tested the influence of stop words (SW) and stemming. We have not tested the influence of the surface syntactic parser, because the parsing was not available the time we have made these tests. The test performed here is done on the Finnish collection with 2003 queries. As the best results is of course, obtained with stop word and stemming, we have then tested the influence of the c constant in order to find out when we reach the optimum. The treatment we apply to both documents and queries is given by:

```
xmlFilterTag | xml2Latin1 | xmldeldia | xmlcase
             | xmlAntiDico -dico common_word.fi
             | xmlcase -noAcc
             | xmlStemFi
```

The first step is filtering the relevant tags from documents or queries. Then we transform XML special characters to their ISO counterpart. We delete all diacritic characters, and change to lower case. At this stage we still have special Finnish characters and accents. We eliminate common words using a list provided by Savoy[1] and then suppress all accents from characters. We apply a Finnish stemmer also proposed by Savoy and modifies to accept XML input/output to produce the final vector. For the queries, we have used the following fields: FI-title FI-desc FI-narr. For documents only the text field has been used.

Results of DFR test with nnn query weighting scheme is in the table 5. When c is zero, then the equation becomes (4), where term weight are all equal for all documents.

run	nRn nnn c=2	ret_rel (483)
raw	29.89	388
SW	35.39	429
stem SW	39.26	452

Table 4: Test weighting nRn nnn

When we examine DFR formula, one can see that when a term does not appear in document d, then only $d_{i,j}$ is null. Then $d^*_{i,j}$ is also equal to zero. If we strictly apply the formula in that case, the weight of the term is still not null and is equal to the formula (4). For practical reason, we have replaced this residual value by zero. This approximation reduces the size of the inverse file, because we do not store null values in file. In fact we have applied the following weighting:

$$w^{'}_{i,j} = \begin{cases} w_{i,j} & \text{if} \quad d_{i,j} \neq 0 \\ 0 & \text{if} \quad d_{i,j} = 0 \end{cases} \tag{3}$$

Table 5 show results for some variation in the constant c.

We can notice that optimum value is about $c = 0.84$. This optimization gain 1.21 points referring neutral value $c = 1$. One can also notice that we obtain more documents in the first 1000 answer for $c = 2$, but the average precision is lower, which means that they are not well sorted.

$$w_{t,d} = \log_2(1 + \lambda_t).\frac{f_t + 1}{n_t} \tag{4}$$

The conclusion of the use of this weighting is that a good constant c value seems to be 0.83. In the rest of the test, we will use the approximated value $c = 0.8$.

For the Okapi weighting, we have use the same value as in [1], that is $k_1 = 1.2$, and $b = 0.75$. In table 9, we have also tested some other value for the French collection: it seems these values are on average good ones.

2.3.1 Testing query weighting

We have tested all combination of the following weight:

[1] http://www.unine.ch/info/clef

c	precision	ret_rel (483)
0.00	4.89	286
0.10	30.24	436
0.50	39.63	448
0.70	40.40	448
0.75	40.90	449
0.80	40.97	449
0.81	41.04	449
0.82	41.06	449
0.83	41.07	449
0.84	41.07	449
0.85	41.02	449
0.86	41.01	449
0.87	41.02	450
0.90	40.16	450
0.95	39.98	450
1.00	39.86	450
1.50	39.41	451
2.00	39.26	452
5.00	39.03	449
10.0	37.96	447

Table 5: Variation of c for nRn nnn (stem AD)

nnn: Only the term frequency is used.

bnn: This is the binary model. Terms presents are associated to the value 1, and 0 otherwise.

lnc: The cosine is the finale normalization. When both used in document and queries, it ensure true vector space model matching, ie. only angle between query et document vector is used. This weighting suppose a log distribution of frequency.

ntc: This is the classical tf*idf measure. We used with queries, the idf is taken from the document collection, not the query collection.

ltc: The same classical measure using log on term frequency.

ltn: The log tf*idf without the cosine normalization.

atn: Normalization with the local maximum term frequency is used with idf.

dtn: The double natural log is used in place of the simple one in ltn.

npn: It is the idf variant use for the Okapi system.

nRn: This is the name for Derivation from randomness.

nOn: This is the name for the Okapi probabilistic weighting.

Results are sum up in the table 6.

We notice that the derivation from randomness model is very stable again the query weighting and that it has the best results in the majority of query weighting. We have the decided to use it for CLEF 2004 in all tests.

	query weighting								
Doc.	nnn	bnn	lnc	ntc	ltc	ltn	atn	dtn	npn
nnn	13.16	9.80	12.22	19.54	19.55	19.55	19.44	19.16	19.82
bnn	28.64	16.61	25.54	34.30	33.67	33.67	33.94	32.50	34.41
atn	26.77	22.65	25.87	28.35	28.02	28.02	28.11	27.85	28.31
ntc	25.72	26.38	25.95	29.26	29.39	29.39	29.60	29.57	29.25
lnc	29.57	23.88	29.75	34.06	35.35	35.35	35.38	25.44	33.99
ltc	32.22	27.84	32.22	32.63	33.00	33.00	32.90	32.44	32.63
ltn	37.71	32.37	37.91	35.99	37.85	37.85	37.86	37.65	36.01
nRn	**41.07**	**36.99**	**40.08**	40.02	**41.29**	**41.29**	**41.05**	**41.92**	40.00
nOn	37.16	29.35	35.95	**40.39**	40.12	40.12	40.32	40.68	**40.12**

Table 6: Query weighting (stem SW c= 0.83)

2.4 French IR

For training, we have used the French corpus of CLEF 2003. We have used our own stemmer, and our own list for removal of common French terms. In this collection, there are 3 sets of documents. For each collection we have selected the following fields: lemonde94 TITLE TEXT, and TI KW LD TX ST for sda 94 and 95. For the queries, we have selected the fields FR-title FR-desc FR-narr. We have tested the same combination of weighting schemes as the one tested in the Finnish collection. The results are in the table 7 and 8.

	Query weighting			
Doc.	nnn	bnn	lnn	ntn
nnn	7.72	2.78	5.71	16.71
bnn	16.01	4.25	13.19	29.73
atn	31.02	27.03	31.16	29.91
ntc	33.53	34.68	35.86	32.09
lnc	36.20	32.22	36.74	39.06
ltc	35.39	35.37	37.40	34.38
ltn	35.65	22.36	32.68	37.87
nRn	**46.98**	**38.15**	**45.01**	**49.06**
nOn	42.25	33.02	40.39	49.01

Table 7: French average precision 1

Finally, we have taken the best weighting query scheme for the Okapi model (nOn) and we have computed some variation of the two constant k_1 and b. The results are in the table 9. The best values are obtained with the couple $(1, 0.75)$ which confirm the choice usually taken for this measure.

In this language, we also demonstrate the stability of the DFR measure (nRn) which performs better than other query weightings, except with binary queries (bnn). We obtain the best average precision with the inverse document frequency (ntn).

We have not performed any special treatments for the queries, like removing terms that are not related to the theme (ex: document, retrieved, etc). The results show that a natural language analysis of the query to remove these empty words should improve the results.

3 Monolingual results

In this part, we comment the results we have obtained at CLEF 2004. We have participated to the monolingual track on French, Finnish and Russian. As we promote the use of syntactic parsing, we

	Query weighting			
Doc.	ltn	atn	dtn	npn
nnn	15.86	15.53	14.47	17.49
bnn	25.13	24.97	23.30	29.15
atn	29.76	30.28	29.47	29.95
ntc	33.89	33.99	33.08	31.98
lnc	40.69	40.82	39.37	38.77
ltc	34.17	34.29	34.73	33.40
ltn	36.64	36.99	35.44	37.89
nRn	**48.16**	**48.76**	**47.03**	**48.78**
nOn	47.07	47.36	45.65	48.38

Table 8: French average precision 2

	b				
k_1	0.25	0.5	0.75	1	1.25
0.5	42.83	45.83	47.04	46.95	46.43
1	46.01	47.96	**49.48**	47.86	44.67
1.5	46.95	48.69	49.36	45.08	41.92
2	46.97	48.56	49.01	43.98	39.04
2.5	46.76	48.19	46.31	43.18	11.81

Table 9: k_1 and b variation for nOn ntn

have submitted mono lingual run all using surface syntactic parsing. Because of time constrains, we have not trained the system with the parsed collection. So we can only compare with CLEF 2003 without natural parsing.

After each parsing, we have transformed the output into a common XML simplified format. One of the main interests in using a natural language parser is the correct normalization of words, the correct detection of compound nouns and correct filtering using lexical categories. For all run, we have choose the derivation from randomness weighting with the constant value fixed to $c = 0.8$, according to the training experiments. No special treatments are done on queries.

These characteristics are important for language which has a morphological derivation like French and Finnish. For all language, we have filtered only nouns, proper noun, verb and adjective. For French, we have used the XIP system from XEROX. After this filtering, we still remove some terms using a stop list, and used also a French stemming. The French queries are weighted using *ntn*. Hence we only modify the weight according to the inverse document frequency. This computation is of course performed using document corpus. The average precision is 44%, which is not an absolute good result. This value is a little lower than our training.

When we examine more closely the results, we discover a big discrepancy between queries. Figure 1, shows the histogram repartition of the 29 queries (from 201 to 250 without 227). There are a lot of query that are either very low precision level (18 queries under 20%) of very high (13 upper 80%).

For the Finnish monolingual run, we obtain an average prevision of 53%, witch is better than the results obtained on CLEF 2003. The histogram in figure 2 shows that 10 query are above 90%, in fact exactly 5 queries reach 100% of precision.

We have use also a surface syntactic parser for the Russian collection, but we cannot compare yet with a more simple raw term indexing because we do not have a Russian stemmer and stop list. The average result of 35% is the lowest for all three languages. Query precision repartition in figure 3 shows that a lot of query (12) have very low precision (under 10%).

The conclusion that we draw is the good behavior of DFR weighting, and probably the benefit of using a surface syntactic parsing on Finnish. In this language, the parser is able to "unglue"

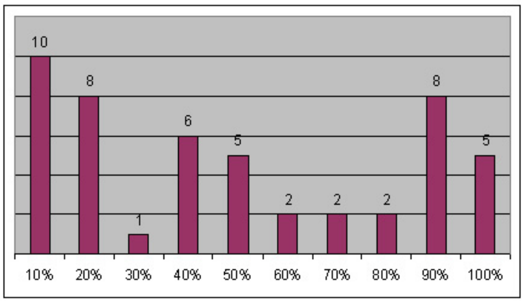

Figure 1: Mono lingual French precision histogram

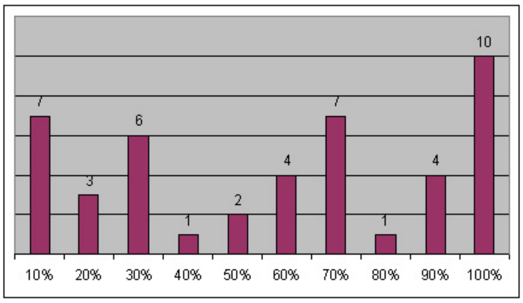

Figure 2: Mono lingual Finnish precision histogram

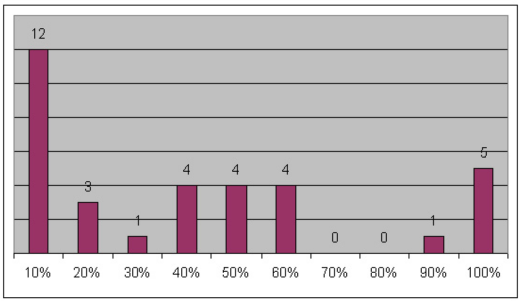

Figure 3: Mono lingual Russian precision histogram

terms and so could achieve better results. We cannot investigate more our results, because we should compare on the same collection the use of the syntactic surface parsing.

4 Topic translation

Bilingual results are obtained by translating the topics using general dictionaries built by compiling several bilingual dictionaries available online (see section 4.1). Then, we experimented 2 methods of translation (see sections 4.2). Both methods take the topic vectors as input and outputs a new translated topic vector.

4.1 Construction of the dictionaries

We compiled 6 bilingual translation dictionaries (see figure 4) using several resources available in house or from Internet. Each resulting dictionary associates a word form to a set of translations and is stored as an XML file (see figure 5).

Dictionary	nb of entry	av. nb of translations per entry	max nb of translations per entry
fr - en	21417	1.92417	22
fr - fi	791	1.06574	4
fr - ru	604	1.06126	3
en - fr	24542	1.67916	25
en - fi	867	1.11649	5
en - ru	15331	2.09901	30

Figure 4: Size of the resulting compiled dictionaries

These dictionaries where compiled from the following sources:

- the Bilingual French-English dictionary from the university of Rennes 1, freely available at `http://sun-recomgen.med.univ-rennes1.fr/Dico/`,
- the FeM dictionary (French English Malay), freely accessible at `http://www-clips.imag.fr/cgi-bin/geta/fem/fem.pl?lang=fr`,
- the French English dictionary available for the participants on the CLEF web site,
- dictionary entries from the Logos website[2],
- the "engrus" English Russian dictionary available on many web sites[3].

As for the French-Russian, French-Finnish and English-Finnish dictionaries, the only available online resource we used is the Logos web site. As it is the only online service we used (other data was available off-line), we chose to only extract entries that were present in the topics to be translated in order to avoid high loads on a public web site. This explains the very small size of these dictionaries.

As French and English were our topic languages of choices, we also reverted the merged French-English dictionaries.

4.2 Topic translation

For each bilingual task we participated in, we propose 2 methods of translation. Both methods take the topic vectors as input and outputs a new translated topic vector.

[2] http://www.logos.it/

[3] see list of mirrors at http://sinyagin.pp.ru/engrus-mirrors.html

English to Russian	English to French

<entry id="issue">
  <equiv>выпускать</equiv>
  <equiv>выпуск</equiv>
  <equiv>излияние</equiv>
  <equiv>результат</equiv>
  <equiv>исход</equiv>
  <equiv>спорный_вопрос</equiv>
  <equiv>исходить</equiv>
  <equiv>вытекать</equiv>
  <equiv>вытекание</equiv>
  <equiv>издать</equiv>
  <equiv>издание</equiv>
  <equiv>происходить</equiv>
  <equiv>vypuskat'</equiv>
  <equiv>выход</equiv>
  <equiv>потомство</equiv>
</entry>
<entry id="harp">
  <equiv>арфа</equiv>
  <equiv>играть_на_арфе</equiv>
</entry>
	<entry id="issue">
	  <equiv>numéro</equiv>
	  <equiv>émettre</equiv>
	  <equiv>émission</equiv>
	  <equiv>matière</equiv>
	  <equiv>délivrer</equiv>
	  <equiv>créer</equiv>
	  <equiv>affliger</equiv>
	  <equiv>impression</equiv>
	  <equiv>question</equiv>
	  <equiv>délivrance</equiv>
	  <equiv>lignée</equiv>
	</entry>
	<entry id="harp">
	  <equiv>harpe</equiv>
	</entry>

Figure 5: Resulting compiled dictionaries sample

4.2.1 Simple topic translation

The first method substitutes each term by all of its available translations. The weight associated to each translation is equal to the weight of the original term divided by the number of available translation (see figure 6).

4.2.2 Filtering by way of an association thesaurus

As one may see in figure 5, many different translations may be found for a single term. Hence, we tried to develop a strategy to give more importance to the "correct" translation(s). For this, we tried to take some context into account, without changing anything to the available lexical resource.

For this, we needed contextual information in each language. Hence, we automatically built an association thesaurus (as exposed in [3]) for each language from the available monolingual documents (see figure 7).

Each association thesaurus is as a graph linking terms. Each arc in the graph links 2 terms that "regularly"[4] appear in the same context. For this experiment, 2 terms are said to be *in the same context* when they appear in the same document.

For our experiment, we assume that terms that are close to each others share some common semantic. We also assume that their "correct" translations should also share the same semantic. Hence, we used these association thesaurus to know if terms and translations share some semantics. Hence, we chose to associate each translations $t_{i,j}$ of a term c_j with a weight $w_{t_{i,j}}$ depending on its distance ($d_{t_{i,j}}$) with the translated context. The distance of a translation to the translated context is given by formula 5.

[4] In this experiment, we filtered out arcs that had a confidence score lower than 20% or higher than 90%.

Original vector (en)	**Translated Vector (ru)**

`<vector id="C250" size="17">`
`<c id="Rabie" w="1"/>`
`<c id="allow" w="1"/>`
`<c id="be" w="1"/>`
`<c id="discussing" w="1"/>`
`<c id="document" w="3"/>`
`<c id="find" w="1"/>`
`<c id="human" w="5"/>`
`<c id="incident" w="2"/>`
`<c id="interest" w="1"/>`
`<c id="learn" w="1"/>`
`<c id="method" w="2"/>`
`<c id="prevention" w="2"/>`
`<c id="rabies" w="4"/>`
`<c id="reader" w="1"/>`
`<c id="relevant" w="1"/>`
`<c id="reporting" w="2"/>`
`<c id="use" w="1"/>`
`</vector>`
	`<vector id="C250" size="57">`
	`...`
	`<!-- Translation of id="reader" w="1" -->`
	`<c id="reader" w="1" untranslated="true"/>`
	`<!-- Translation of id="human" w="5" -->`
	`<c id="ЧЕЛОВЕЧЕСКИЙ" w="2.5"/>`
	`<c id="ЧЕЛОВЕК" w="2.5"/>`
	`<!-- Translation of id="document" w="3" -->`
	`<c id="ПОДТВЕРЖДАТЬ_ДОКУМЕНТАМИ" w="1"/>`
	`<c id="ДОКУМЕНТ" w="1"/>`
	`<c id="СВИДЕТЕЛЬСТВО" w="1"/>`
	`<!-- Translation of id="Rabie" w="1" -->`
	`<c id="Rabie" w="1" untranslated="true"/>`
	`<!-- Translation of id="prevention" w="2" -->`
	`<c id="prevention" w="2" untranslated="true"/>`
	`<!-- Translation of id="interest" w="1" -->`
	`<c id="ИНТЕРЕС" w="0.25"/>`
	`<c id="ВЫГОДА" w="0.25"/>`
	`<c id="ИНТЕРЕСОВАТЬ" w="0.25"/>`
	`<c id="ЗАИНТЕРЕСОВЫВАТЬ" w="0.25"/>`
	`<!-- Translation of id="rabies" w="4" -->`
	`<c id="БЕШЕНСТВО" w="4"/>`
	`...`
	`</vector>`

Figure 6: Simple topic translation

$$\begin{gathered} d_{t_{i,j}} = Min(d(t_{i,j}, t_{k,l}); \forall l, k \mid l \neq j, 1 \leq k \leq [T_l|) \\ \text{where } t_{k,l} \in T_l \\ \text{and } T_l \text{ is the set of translation of the term } c_l \\ \text{and } d(t_{i,j}, t_{k,l}) \text{ is the minimal distance in the target thesaurus between } t_{i,j} \text{ and } t_{k,l} \end{gathered} \tag{5}$$

$$\begin{gathered} w_{t_{i,j}} = \begin{cases} w_j / d_{i,j} & \text{if} \quad d_{i,j} \neq 0 \\ w_j / \lvert T_j \rvert & \text{if} \quad d_{i,j} = 0 \end{cases} \\ \text{where } w_j \text{ is the weight of the source term } c_j \text{ in the source vector} \end{gathered} \tag{6}$$

Figure 8 shows a sample resulting translated vector. One may notice the higher weight of the selected translations (e.g. interest → ИНТЕРЕС).

4.3 Discussion

CLIPS results on the bilingual tasks are rather disappointing, with interpolated recall-precision averages at 0.00 dropping from 57.68% (Monolingual Russian) to 17.1% with simple topic translation (Bilingual English-Russian: CLIPSENRU1) and even to 8.59% with filtered topic translation (CLIPSENRU2).

The main reason for this drop is certainly due to the lack of wide coverage bilingual lexical resources. The dictionaries we used were very small and did not provide translations for many terms of the topics. This is especially true for French to Russian and Finnish lexical resources where 60% to 70% of the source terms are not translated. However, English to Russian lexicon was a little better, and about 18% of the terms remain without translation.

Corpus	nb of term in the corpus	nb of arcs in the thesaurus	nb of terms left in the thesaurus
LeMonde95	134786	21717	4247
GH95	151595	23605	4891
Izvestia95	43346	23992	2466
Aamu95	271860	19138	9000

Figure 7: Size of the association thesaurii

However, this does not explain the drop in interpolated recall-precision averages when filtering the translations through the association thesaurii, as it does not change the set of translations, but only the weight of those translations. Moreover, when manually evaluating the weighted translation, one usually agree with the translation that are chosen.

We think that 2 factors explains theses drops:

- First, in the simple topic translation method, the weight of each translation is divided by the number of translations for the source term. This lowers the relative importance of terms that bear many translations, (which is usually the case of general nouns or support verbs).
- Second, when raising the weight of "correct" translations by way of the association thesaurii, we also raise the weight of such general terms. Hence, we give more importance to terms that do not bear any thematic closeness with the requested documents (and this is especially the case with CLEF topics that are instructions usually containing "find documents reporting on..." or "find information on...").

5 Conclusion

All run are performed on the collection parsed using a syntactic surface parsing. Best monolingual results are obtained for the Finnish collection, probably because of the correct word splitting. We have to redo the tests with no analyzer to have a strong conclusion on its use in an IR context/

Bilingual results are disappointing but they are partly explained by the difficulty in finding wide coverage lexical resources for languages in which we previously had no experience whatsoever.

The filtering of translations through association thesaurii is rather interesting, even if we did not have enough time to use it appropriately. This technique may also be interesting in translation selection tasks or, with adaptation, on lexical disambiguation tasks. It's main interest in such tasks comes from the fact that it does not require any special training data (like parallel documents or manually disambiguated corpora) as association thesaurii may be computed automatically from the corpus. Hence such technique may easily bring some result in those tasks in any language, provided that monolingual data is available as well as an automatic process to lemmatize such corpora.

References

[1] Gianni Amati, Claudio Carpineto, and Giovanni Romano. Comparing weighting models for monolingual information retrieval. In *CLEF 2003*, Trondheim, Norway, 2003.

[2] Gianni Amati and Cornelis Joost van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transaction on Information Systems*, 20(4):357–389, October 2002.

Original vector (en)	Translated Vector (ru)
<vector id="C250" size="17"> <c id="Rabie" w="1"/> <c id="allow" w="1"/> <c id="be" w="1"/> <c id="discussing" w="1"/> <c id="document" w="3"/> <c id="find" w="1"/> <c id="human" w="5"/> <c id="incident" w="2"/> <c id="interest" w="1"/> <c id="learn" w="1"/> <c id="method" w="2"/> <c id="prevention" w="2"/> <c id="rabies" w="4"/> <c id="reader" w="1"/> <c id="relevant" w="1"/> <c id="reporting" w="2"/> <c id="use" w="1"/> </vector>	<vector id="C250" size="57"> ... <!-- Translation of id="reader" w="1" --> <c id="reader" w="1" untranslated="true"/> <!-- Translation of id="human" w="5" --> <c id="ЧЕЛОВЕЧЕСКИЙ" w="2.5"/> <c id="ЧЕЛОВЕК" w="5"/> <!-- Translation of id="document" w="3" --> <c id="ПОДТВЕРЖДАТЬ_ДОКУМЕНТАМИ" w="1"/> <c id="ДОКУМЕНТ" w="1"/> <c id="СВИДЕТЕЛЬСТВО" w="1"/> <!-- Translation of id="Rabie" w="1" --> <c id="Rabie" w="1" untranslated="true"/> <!-- Translation of id="prevention" w="2" --> <c id="prevention" w="2" untranslated="true"/> <!-- Translation of id="interest" w="1" --> <c id="ИНТЕРЕС" w="0.5"/> <c id="ВЫГОДА" w="0.25"/> <c id="ИНТЕРЕСОВАТЬ" w="0.25"/> <c id="ЗАИНТЕРЕСОВЫВАТЬ" w="0.25"/> <!-- Translation of id="rabies" w="4" --> <c id="БЕШЕНСТВО" w="4"/> ... </vector>

Figure 8: Topic translation with filtering

[3] Jean-Pierre Chevallet and Gilles Serrasset. Simple translations of monolingual queries expanded through an association thesaurus. x-iota ir system used for clips bilingual experiments. In *CLEF 2003*, 2003.

[4] S. E. Robertson. Overview of the okapi projects. *Journal of Documentation*, 53(1):3–7, 1997.

[5] Steve E. Robertson, S. Walker, and Micheline Baulieu. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive track. In *Preceedings of TREC-7*, 1998.