A Possibilistic Query Translation Approach for Cross-Language Information Retrieval

Wiem Ben Romdhane[1], Bilel Elayeb[1,2], Ibrahim Bounhas[3], Fabrice Evrard[4], and Narjès Bellamine Ben Saoud[1]

[1] RIADI Research Laboratory, ENSI, Manouba University 2010, Tunisia
br.wiem@yahoo.fr, Narjes.Bellamine@ensi.rnu.tn
[2] Emirates College of Technology, P.O. Box: 41009, Abu Dhabi, United Arab Emirates
Bilel.Elayeb@riadi.rnu.tn
[3] LISI Lab. of computer science for industrial systems, ISD, Manouba University 2010, Tunisia
Bounhas.Ibrahim@yahoo.fr
[4] IRIT-ENSEEIHT, 02 Rue Camichel, 31071 Toulouse Cedex 7, France
Fabrice.Evrard@enseeiht.fr

Abstract. In this paper, we explore several statistical methods to find solutions to the problem of query translation ambiguity. Indeed, we propose and compare a new possibilistic approach for query translation derived from a probabilistic one, by applying a classical probability-possibility transformation of probability distributions, which introduces a certain tolerance in the selection of word translations. Finally, the best words are selected based on a similarity measure. The experiments are performed on CLEF-2003 French-English CLIR collection, which allowed us to test the effectiveness of the possibilistic approach.

Keywords: Cross-Language Information Retrieval (CLIR), Query Translation, Possibilistic Approach.

1 Introduction

With the huge expansion of documents in several languages on the Web and the increasing desire of non-native speakers of the English language to be able to retrieve documents in their own languages, the need for Cross-Language Information Retrieval (CLIR) System has become increasingly important in recent years. In fact, in the CLIR task, either the documents or the queries are translated. However, the majority of approaches focus on query translation, because document translation is computationally expensive. There are three main approaches to CLIR: Dictionary-based methods, parallel or comparable corpora-based methods, and machine translation methods.

The *Dictionary-based methods* [16][14] are the general approaches for CLIR when no commercial MT system with a recognized reputation is available. Several information retrieval systems (IRS) have used the so-called "bag-of-words" architectures, in which documents and queries are decayed into a set of words (or phrases) during an indexing

D.-S. Huang et al. (Eds.): ICIC 2013, LNAI 7996, pp. 73–82, 2013.

procedure. Therefore, queries can be simply translated by replacing every query term with its corresponding translations existing in a bilingual term list or a bilingual dictionary. Nevertheless, dictionary-based methods suffer from several difficulties such as: i) no translation of non-existing specific words in the used dictionary; ii) the addition of irrelevant information caused by the intrinsically ambiguities of the dictionary; iii) the decreasing of the effectiveness due to the disappointment to translate multiword expressions. To reduce ambiguity, one may adopt a corpus-based approach.

In *corpus-based methods* [17], a set of multilingual terms extracted from parallel or comparable corpora is exploited. Approaches based statistical/probabilistic method on parallel text written in multiple languages with the intention of selecting the correct word translation provides a good performance, but they suffer from many drawbacks. Firstly, the translation association created among the parallel words in the text is generally domain restricted, which means that accuracy decreases outside the domain. Secondly, parallel texts in different pairs of languages, are not always available.

In *machine translation (MT) techniques* [5][13], the main aim is to analyze the context of the query before translating its words. In fact, syntactic and semantic ambiguities are the principal problems decreasing MT performance. Besides, MT-based approaches suffer from several others limits decreasing the effectiveness of CLIR. Firstly, MT systems have serious difficulties to appropriately generate the syntactic and semantic analysis of the source text. Secondly, full linguistic analysis is computationally expensive, which decreases search performance.

In fact, query translation approaches need training and matching models which compute the similarities (or the relevance) between words and their translations. Existing models for query translation in CLIR are based on poor, uncertain and imprecise data. While probabilistic models are unable to deal with such type of data, possibility theory applies naturally to this kind of problems [8]. Thus, we propose a possibilistic approach for query translation derived from a probabilistic one using a probability/possibility transformation [6]. This approach begins with a query analysis step, then a lexical analysis step, and finally the selection of the best translation using different similarity measures.

This paper is organized as follows. Section 2 details our approach which is experimented in section 3. In section 4, we conclude our work and give some directions for future research.

2 The Proposed Approach

We propose a new possibilistic approach for query translation in CLIR. The proposed approach is an extension of a probabilistic model proposed by [12] into a possibilistic framework, using an existing probability/possibility transformation method [6]. In this approach we used a greedy algorithm to choose the best translation [12]. The calculation of similarity between the terms and the cohesion of a term x with a set X of other terms are two essential steps before selecting the best term translation. In our case, we used the *EMMI* weighting measure [15] to estimate the probabilistic similarity between terms. Then, we extended it to a possibilistic framework (*EMMI-POSS*) using

an existing probability/possibility transformation [6]. We briefly recall in the following this transformation and we detail the three main steps of our query translation process, which are summarized in figure 1.

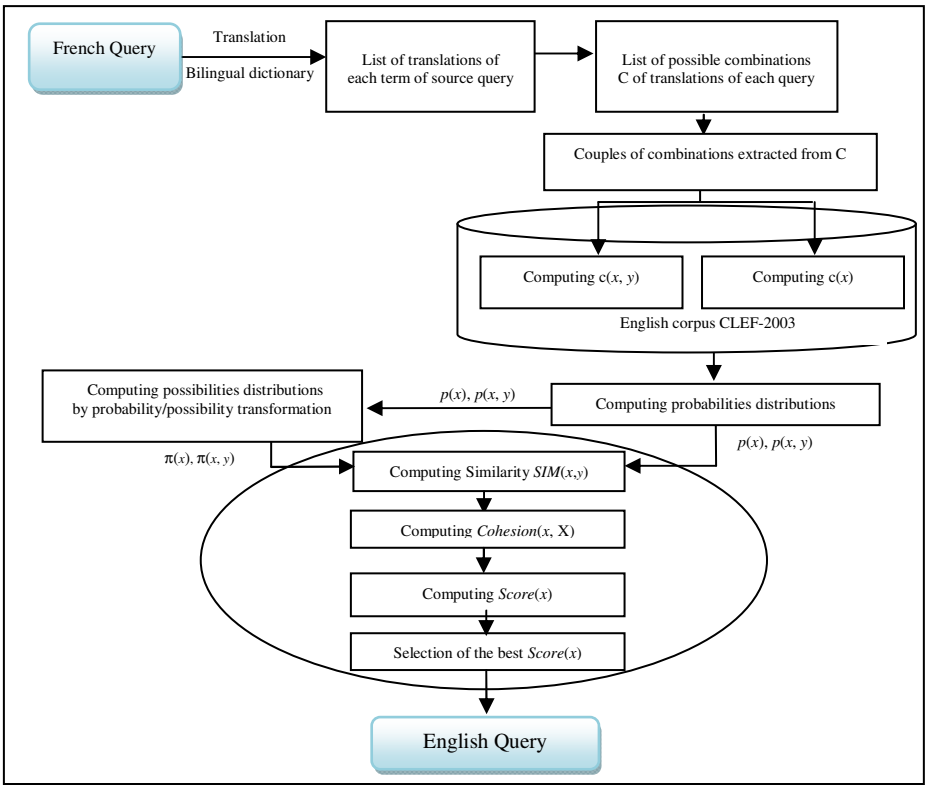

Fig. 1. Overview of query translation process

Formally the similarity between the terms x and y is given by formula (1). However, the cohesion of a term x with a set X of other words is the maximum similarity of this term with each term of the set, as given by formula (4).

$$SIM(x, y) = p(x, y) \times \log_2\left(\frac{p(x, y)}{p(x) \times p(y)}\right) \quad (1)$$

$$p(x, y) = \frac{c(x, y)}{c(x)} + \frac{c(x, y)}{c(y)} \quad (2)$$

$$p(x) = \frac{c(x)}{\sum_x c(x)}; \; p(y) = \frac{c(y)}{\sum_y c(y)} \quad (3)$$

$$Cohesion(x, X) = \max_{y \in X}(SIM(x, y)) \quad (4)$$

Where $c(x, y)$ is the frequency that the term x and the term y co-occur in the same sentences in the collection. $c(x)$ is the number of occurrences of term x in the collection.

2.1 Probability/Possibility Transformation

Given the universe of discourse $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ and a probability distribution p on Ω, such that $p(\omega_1) \geq p(\omega_2) \geq \ldots \geq p(\omega_n)$, we can transform p into a possibility distribution π using the following formulas (for more detail you can see [6][7]):

$$\pi(\omega_i) = i * p(\omega_i) + \sum_{j=i+1}^{n} p(\omega_j), \forall i = 1, \ldots, n \quad (5)$$

$$\sum_{j=1}^{n} p(\omega_j) = 1 \quad \text{and} \quad p(\omega_{n+1}) = 0 \text{ by convention.} \quad (6)$$

Among several transformation formulas, we have chosen this formula, because it satisfies both the probability/possibility consistency (i.e. $\Pi(A) \geq P(A)$) and the preference preservation principles (i.e. $p(\omega_i) > p(\omega_j) \Leftrightarrow \pi(\omega_i) > \pi(\omega_j)$ [7]). Indeed, this transformation process has allowed us to increase the possibilistic scores of coexistence of two terms in order to penalize the scores of terms that are weakly co-occurring. In fact, the penalty and the increase of scores are proportional to the power of words to discriminate between the possible combinations of coexistence.

Example: Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and a probability distribution p on Ω such that: $p(\omega_1)=0.2$; $p(\omega_2)=0.5$; $p(\omega_3)=0.3$; $p(\omega_4)=0$. So, we have: $p(\omega_2) > p(\omega_3) > p(\omega_1) > p(\omega_4)$. By applying the transformation formula, we have: $\pi(\omega_2)=(1*0.5)+(0.3+0,2)=1$; $\pi(\omega_1) = (3*0.2) + 0 = 0.6$; $\pi(\omega_3) = (2*0.3) + 0.2 = 0.8$; $\pi(\omega_4) = (4*0) + 0 = 0$.

2.2 Query Analysis

It is the first step in this approach, in which stop words are deleted from the source queries using a list of words considered as non-significant for source queries. Then, we extract the set of possible translations from a French-English dictionary generated using the free online dictionary Reverso[1].

2.3 Lexical Analysis

The step of lemmatization aims to find the canonical form of a word, so that different grammatical forms or variations are considered as instances of the same word. We applied the process of lemmatization on the test collection and queries before their translations. This reduction mechanism gives better results for the following matching phase.

2.4 Selection of Best Translation

It is the main step in our approach. Indeed, selecting the best translation among several ones existing in the bilingual dictionary is summarized as follows. The suitable translations of source query terms co-occur in the target language documents contrary to incorrect translations one. Consequently, we select for each set of the source query

[1] `http://www.reverso.net/text_translation.aspx?lang=FR`

terms the best translation term, which frequently co-occurs with other translation terms in the target language. However, it is computationally very costly to identify such an optimal set. For that reason, we take advantage from an approximate Greedy algorithm as used in [12].We briefly summarize in the following the main principle of this algorithm. Firstly, and using the bilingual dictionary, we select a set T_i of translation terms for each of the n source query terms $\{f_1,\ldots,f_n\}$. Secondly, we compute the cohesion of every term in each set T_i with the other sets of translation terms. The best translation in each T_i has the maximum degree of cohesion. Finally, the target query $\{e_1,\ldots,e_n\}$ is composed of the best terms from every translation set.

Cohesion is based on the similarity between the terms. We transform the weighting measure *EMMI* to a possibilistic one (*EMMI-POSS*), which is successfully used to estimate similarity among terms. However, the measure *EMMI-POSS* does not take into account the distance between words. In fact, we observe that the local context is more important for the selection of translation. If two words appear in the same document, but in two remote locations, it is unlikely to be strongly dependent. Therefore, a distance factor was added by [12] in computing word similarity.

2.5 Illustrative Example

Let us consider the following French source query Q: {*L'Union Européenne et les Pays Baltes*}. Indeed, we have a set of possible translations for each term in the query Q from the used dictionary. The term "*union*" has two possible translations (*union, unity*), the term "*Européenne*" has the unique translation (*european*), the term "*pays*" has two possible translations (*country, land*) and the term "*Baltes*" has the unique translation (*Baltic*). In fact, Given the source query Q, we generate the set of possible translations combinations from a bilingual dictionary. In this example, there are 4 possible translation combinations (cf. table 1). The best translation is which has the greater possibilistic score. Table 2 and give detail of calculus.

Table 1. Translation combinations for {*L'Union Européenne et les Pays Baltes*}

	Translation Combinations
1	union AND european AND country AND Baltic
2	union AND european AND land AND Baltic
3	unity AND european AND country AND Baltic
4	unity AND european AND land AND Baltic

Probability values in table 2 and 3 are very low comparing to the possibility ones. So, they have a poor discriminative effect in the selection of the suitable translation. Consequently, we risk having very close probabilistic similarity scores, in which ambiguity translation cannot be correctly resolved. Moreover, the selected English translation of the given French source query {union, européenne, pays, balte} is the target Enlish query {unity, european, country, baltic}. We remark here that the suitable translation of the name phrase (NP) "union européenne" is not "European unity" but "European union". Consequently, we mainly need to identify the NP in the source query and translate them before translating one-word terms.

Table 2. Possibilistic similarity scores for the different pairs of words (x, y)

Pairs of words (x, y)	C(x, y)	P(x, y)	π(x, y)	SIM(x, y)
union-european	1317	0.2642	9.3428	34.0590
union - country	209	0.0442	4.9091	12.9894
union - baltic	6	0.0583	5.5669	43.4782
european - country	535	0.0894	6.6168	20.0428
european -baltic	5	0.0485	5.1314	39.2299
country –baltic	8	0.0872	6.5532	51.9453
union – european	1317	0.2642	9.3428	34.0590
union - land	1	0.0077	1.5429	2.2342
union - baltic	6	0.0583	5.5669	43.4782
european-land	51	0.0134	2.3534	4.7296
european -baltic	5	0.0485	5.1314	39.2299
land - baltic	1	0.0097	1.8719	12.3381
unity - european	15	0.0380	4.5408	25.4750
unity - country	10	0.0282	3.8353	20.3097
unity -baltic	0	0.0	0.0	0.0
european - country	535	0.0894	6.6168	20.0428
european - baltic	5	0.0485	5.1314	39.2299
country - baltic	8	0.0872	6.5532	51.9453
unity -european	15	0.0380	4.5408	25.4750
unity - land	1	0.0024	0.5520	1.6403
unity-baltic	0	0.0	0.0	0.0
european-land	51	0.0134	2.3534	4.7296
european - baltic	5	0.0485	5.1314	39.2299
land - baltic	1	0.0097	1.8719	12.3381

Table 3. The final possibilistic score of each possible translation

Possible translations of word x	C(x)	P(x)	π(x)	Score (x)
union	9894	0.0148	0.8499	14.0363
unity	455	0.0006	0.1058	**50.9500**
european	11121	0.0165	0.8784	**119.0682**
country	13469	0.0204	0.9228	**103.8907**
land	5592	0.0083	0.6653	24.6762
baltic	108	0.0001	0.0291	**186.5989**

3 Experimental Evaluation

Our experiments are performed through our possibilistic information Retrieval System [10], and implemented using the platform Terrier[2]. It provides many existing matching models such as OKAPI and a new possibilistic matching model proposed by [11]. We propose and compare here our results using these two matching model in order to study the generic character of our approach.

[2] http://terrier.org/

Experiments are achieved using a subset of the collection CLEF-2003. This part includes articles published during 1995 in the newspaper "*Glasgow Herald*". This collection consists of 56472 documents and 54 queries, forming 154 MB. We only take into account the part <title> of the test queries, because it contains several isolate words, which are suitable to experiment our approach. However, we plan to consider other part of queries such as <description> and <narrative>, in which the context is relevant in the translation process. To evaluate our possibilistic approach, we compare our results to some existing probabilistic similarity measures such as T-score (*TS*) [3], Log Likelihood Ratio (*LLR*) score, [9], Dice Factor (*DF*) [16] and Mutual Information (*MI*) [4].

Table 4 contains statistics on two elements u and v which are in this case, the components of an expression. O_{11} is the number of co-occurrences of u with v. O_{12} is the number of occurrences of u with an element other than v, etc.

Table 4. The contingency table

	$t_1 = v$	$t_1 \neq v$
$t_2 = u$	O_{11}	O_{12}
$t_2 \neq u$	O_{21}	O_{22}

We have also:

$$R_1 = O_{11} + O_{12} \tag{7}$$

$$R_2 = O_{21} + O_{22} \tag{8}$$

$$C_1 = O_{11} + O_{21} \tag{9}$$

$$C_2 = O_{12} + O_{22} \tag{10}$$

$$N = R_1 + R_2 = C_1 + C_2. \tag{11}$$

We also calculate the expected frequency of collocation as follows:

$$E_{11}=(R_1*C_1)/N \tag{12}$$

The *LLR,MI, TS* and *DF* score are calculated as follows:

$$\mathrm{LLR(u,\ v)} = -2\ \log\left(\frac{\mathrm{L(O_{11}, C_1, r)} * \mathrm{L(O_{12}, C_2, r)}}{\mathrm{L(O_{11}, C_1, r_1)} * \mathrm{L(O_{12}, C_2, r_2)}}\right) \tag{13}$$

$$\mathrm{L}(k,n,r) = k^r * (1-r)^{n-k} \ where: r = R_1 / N, r_1 = O_{11} / C_1, r_2 = O_{12} / C_2 \tag{14}$$

$$MI = \log_2(O_{11} / E_{11}) \tag{15}$$

$$TS = \frac{(O_{11} - E_{11})}{\sqrt{O_{11}}} \tag{16}$$

$$DF = 2 * \frac{O_{11}}{R_1 + C_1} \tag{17}$$

The proposed approach is assessed using the mean average precision (MAP) as a performance measure. The formula of computing the MAP is the following:

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(doc_i) \tag{18}$$

Where:
Q_j : The number of relevant documents for query j;
N: The number of queries;
$P(doc_i)$: The precision at the i^{th} relevant document.

Moreover, we compare the possibilistic approach (*EMMI-POSS*) both to monolingual IR task and to others probabilistic similarity measures, using OKAPI and Possibilistic matching models (Figure 2 and 3, respectively). In fact, we used the precision (Y-axe) over 11 points of recall in the X-axe (0.0, 0.1, ..., 1.0) to draw all recall-precision curves.

Using OKAPI (figure 2) or the possibilistic (figure 3) matching model, results in both figures show that the possibilistic query translation approach has the closest recall-precision curve to the Monolingual task, which confirm its effectiveness comparing to other probabilistic approaches. Indeed, the discriminative character of the possibilistic approach improves its ability to solve the problem of query translation ambiguity and consequently enhance its efficiency.

On the other hand, the mean average precision of *EMMI-POSS* (0.23) is very close to that obtained for the Monolingual (0.24) and that obtained for *TS* (0.21). The *LLR* metric has the worst result (0.15). These results are also confirmed using the possibilistic matching model. Indeed, results in figure 3 prove that the mean average precision of *EMMI-POSS* (0.165) is very close to that obtained for the Monolingual (0.17). The *LLR* metric stays also the worst one with 0.104.

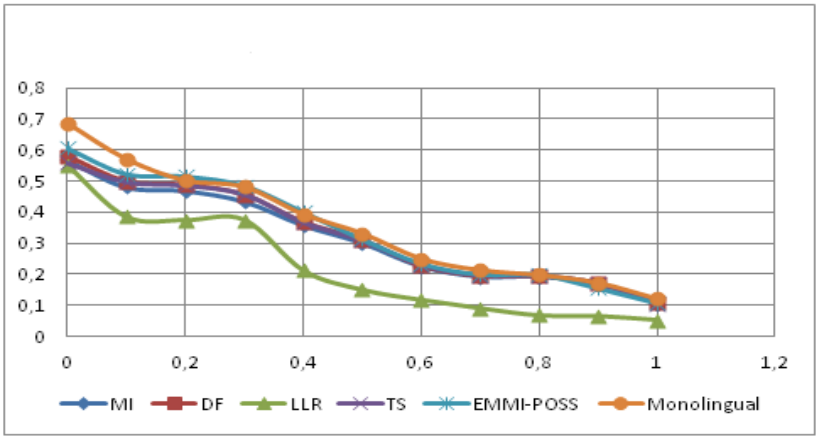

Fig. 2. Recall-Precision curves of Monolingual vs. All similarity measures (OKAPI)

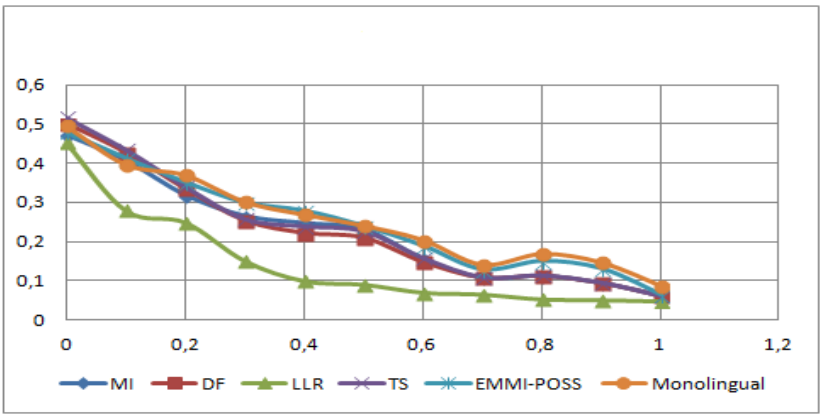

Fig. 3. Recall-Precision curves of Monolingual vs. All similarity measures (Possibilistic)

In fact our approach for CLIR has some drawbacks such us: (i) the limited coverage of dictionary and; (ii) The complexity of the algorithm allowing to choose the suitable translation among the set of the translations proposed by the dictionary. To overcome these limitations, we exploited the cohesion between a given query term and its possible translations in the training corpus and a particular similarity score measure to select the suitable translation of each query term. However, the results were mainly influenced by the specific properties of the used document collection.

4 Conclusion

In this paper, we presented a possibilistic query translation approach based on the cohesion between the translations of words. This approach is based on probability/possibility transformation improving discrimination in the selection of suitable translation. Besides, this transformation did not increase the complexity such as in [1][2]. We have tested and compared several similarity scores to improve query translation based dictionaries in CLIR.

The idea of applying possibility theory to query translation is identical to the use of probabilities in the Bayesian probability model. In fact, it is necessary to evaluate many parameters, a task that cannot be compatible with poor data. The problem of accurately estimating probability distributions for probabilistic query translation is important for the accurate calculation of the probability distribution of translations. However, due to the use of the product to combine probability values (which are frequently small), the probability estimation error may have a significant effect on the final estimation. This contrasts with the possibility distributions which are less sensitive to imprecise estimation for several reasons. Indeed, a possibility distribution can be considered representative of a family of probability distributions corresponding to imprecise probabilities, which are more reasonable in the case of insufficient data (such as the case when some words do not exist in the bilingual dictionary).

Furthermore, we no longer need to assume a particular form of probability distribution in this possibilistic reconciliation process.

References

1. Bounhas, M., Mellouli, K., Prade, H., Serrurier, M.: Possibilistic classifiers for numerical data. Soft Computing 17, 733–751 (2013)
2. Bounhas, M., Mellouli, K., Prade, H., Serrurier, M.: From Bayesian Classifiers to Possibilistic Classifiers for Numerical Data. In: Deshpande, A., Hunter, A. (eds.) SUM 2010. LNCS, vol. 6379, pp. 112–125. Springer, Heidelberg (2010)
3. Church, K., Gale, W., Hanks, P., Hindle, D.: Using statistics in lexical analysis. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, pp. 115–164. Lawrence Erlbaum Associates, Hillsdale (1991)
4. Daille, B.: Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques. Ph.D. Thesis, University of Paris 7 (1994) (in French)
5. Mavaluru, D., Shriram, R., Banu, W.A.: Ensemble Approach for Cross Language Information Retrieval. In: Gelbukh, A. (ed.) CICLing 2012, Part II. LNCS, vol. 7182, pp. 274–285. Springer, Heidelberg (2012)
6. Dubois, D., Prade, H.: Unfair coins and necessity measures: Towards a possibilistic interpretation of histograms. Fuzzy Sets and Systems 10, 15–20 (1985)
7. Dubois, D., Prade, H., Sandri, S.: On Possibility/Probability transformation. Fuzzy Logic: State of the Art, 103–112 (1993)
8. Dubois, D., Prade, H.: Possibility Theory: An Approach to computerized Processing of Uncertainty. Plenum Press, New York (1994)
9. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19, 61–74 (1994)
10. Elayeb, B., Bounhas, I., Ben Khiroun, O., Evrard, F., Bellamine Ben Saoud, N.: Towards a Possibilistic Information Retrieval System Using Semantic Query Expansion. International Journal of Intelligent Information Technologies 7, 1–25 (2011)
11. Elayeb, B., Evrard, F., Zaghdoud, M., Ben Ahmed, M.: Towards an Intelligent Possibilistic Web Information Retrieval using Multiagent System. The Interactive Technology and Smart Education, Special issue: New learning support systems 6, 40–59 (2009)
12. Gao, J., Nie, J.Y., Xun, E., Zhang, J., Zhou, M., Huang, C.: Improving Query Translation for Cross-Language Information Retrieval using Statistical Models. In: Proceedings of SIGIR 2001, New Orleans, Louisiana, USA, pp. 9–12 (2001)
13. Iswarya, P., Radha, V.: Cross Language Text Retrieval: A Review. International Journal Of Engineering Research and Applications 2, 1036–1043 (2012)
14. Mallamma, V.R., Hanumanthappa, M.: Dictionary Based Word Translation in CLIR Using Cohesion Method. In: INDIACom-(2012) ISSN 0973-7529, ISBN 978-93-80544-03-8
15. Rijsbergen, V.: Information Retrieval. Butterworths, Londres (1979)
16. Smadja, F., Mckeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: a statistical approach. Computational Linguistics 22, 1–38 (1996)
17. Vitaly, K., Yannis, H.: Accurate Query Translation For Japanese-English Cross-Language Information Retrieval. In: International Conference on Pervasive and Embedded Computing and Communication Systems, pp. 214–219 (2012)