Improving Query Translation with Confidence Estimation for Cross Language Information Retrieval

Youssef Kadri
Lab. RALI, DIRO, Université de Montréal,
Montreal, Quebec, H3C 3J7 Canada
kadriyou@iro.umontreal.ca

Jian-Yun Nie
Lab. RALI, DIRO, Université de Montréal, Canada
Montreal, Quebec, H3C 3J7 Canada
nie@iro.umontreal.ca

Categories and Subject Descriptors
H.3.3 [Information storage and retrieval]: Information Search and Retrieval – *retrieval models, query formulation.*

General Terms
Algorithms, Performance, Experimentation, Theory.

Keywords
CLIR, confidence estimation.

1. INTRODUCTION
Query translation is the key problem in Cross Language Information Retrieval (CLIR). This can be done by using Bilingual Dictionaries (BDs), parallel corpora or machine translation [4]. When several translation tools or resources are combined, a crucial problem is to combine and re-weight all the translation candidates correctly. In the previous studies, simple methods are usually employed, i.e. one combines various translations for the same query term linearly by assigning or optimizing automatically a confidence weight to the translation tool or resource [3, 5]. However, we notice that a single confidence score is assigned to all the translations from the same translation resource. This score does not modify the relative importance of the translations from the same resource. In practice, sometimes when new criteria are considered, a translation with a low score suggested by a resource can turn out to be a better translation. In this paper, we propose a confidence estimation technique to adjust the weight of various possible translations of a query term. Given different translations from different resources, our approach estimates a confidence for each translation candidate of a query term according to additional informative features. According to these confidence measures, the translation candidates are re-weighted and homogenous weights are assigned to the translation candidates from different resources. Therefore, the advantages of this approach are twofold. On one hand, the confidence measure allows us to adjust the original weight of the translations and to select the best translation terms. On the other hand, the confidence estimates also provide us with a comparable weighting for the translation candidates across different translation resources.

2. CONFIDENCE ESTIMATION
Confidence estimation was originally used in speech recognition and understanding (Hazen and al, 2002). It has been applied to improve recognition by incorporating extra information into the recognition process. In CLIR, query translation is performed with many resources for the same word. We then have the same problem as in speech recognition: selecting the correct translation(s) among all the candidates. In this context, confidence measures can be used to learn how to adjust the original scores of translations by observing their performance on new texts. These confidence estimates will be used in this paper as a uniform measure on translations instead of the original probabilities. Concretely, for a given translation produced by any resource, STM or BD, we aim to measure the confidence of it being correct, according to some informative features.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011..

3. CONFIDENCE MEASURES IN CLIR
Our work focuses on the estimation of the translation model $p(t_F \mid q_E)$. Traditionally when translation is done with more than one resource, linear combination is used to estimate the translation model as follows:

$$p(t_F \mid q_E) = z_{q_E} \sum_i \lambda_i p_i(t_F \mid q_E) \quad (1)$$

Where λ_i is the parameter related to the translation resource *i* and z_{q_E} is a normalization factor so that $\sum_{t_F} p(t_F \mid q_E) = 1$. The parameters λ denote the confidence weight assigned to each resource. These parameters can be optimized using some training data. $p_i(t_F \mid q_E)$ is the probability of translating the source word q_E to the target word t_F by the resource *i*. As we discussed earlier, this method is unable to re-rank the translation candidates from the same resource. A crucial question that we have to ask is: Given a translation candidate, is it correct and how confident are we on its correctness?

Confidence estimation is used to answer this very question. Therefore, instead of trusting blindly the weights assigned by different resources, we use confidence estimation to reconsider each of the translation candidates according to additional features (such as the POS-tagging, the rank of the candidate, etc.). Given a translation candidate t_F for a source term q_E and *X*— a set of relative features, $p(t_F \mid q_E)$ is computed as the sum of confidence estimates on this candidate using different resources, i.e.

$$p(t_F \mid q_E) = z_{q_E} \sum_i p_i(C = 1 \mid t_F, q_E, X) \quad (2)$$

Where $p(C = 1 \mid t_F, q_E, X)$ is normalized such that: $\sum_{t_F} p(C = 1 \mid t_F, q_E, X) = 1$.

4. COMPUTATION OF CONFIDENCE

Confidence for a translation is defined as the posterior probability that this translation is correct $P(C=1|X)$, given X— the source word, a translation and a set of relative features. We use a Multi Layer Perceptron (MLP) to estimate the probability of correctness $P(C=1|X)$ of a translation. The MLP was trained to minimize the negative log-likelihood (or cross entropy CE) assigned to the test corpus by the model normalized by the number of examples in the test corpus [1].

4.1 Features

We selected intuitively seven classes of features hypothesized to be informative for the correctness of a translation. These features are: index of the translation source, translation probabilities and reverse translation probabilities, rank of the translation, source sentence-related features such as the frequency of the source word in the source sentence, language model features (unigram language model for source and target words on the training data) and Part-Of-Speech (POS) tags of both source word and translation candidate (i.e. lexical tag probabilities).

4.2 Experiments on confidence measure

The corpus used to train confidence is extracted from two parallel corpora: The Hansard corpus and the Web corpus. The former is composed of debates in the Canadian parliament, while the latter is automatically gathered from the Web [2]. It consists of around 60 K pairs of aligned sentences. Source sentences are translated word by word using baseline models (two STMs and a BD). We translated each source word with the five most probable translations for the STMs and all the translations provided by the BD. Translations are then compared to the reference sentence to build a labelled corpus: if a candidate translation appears in the reference translation sentence, then it is considered to be correct.

We test with various numbers of hidden units (from 0 to 100). We used the Normalized Cross Entropy (NCE) to compare the performance of different architectures. The NCE measures the relative drop in negative log-likelihood compared to the baseline that depends on the prior probability of correctness. The MLP with 20 hidden units gave the best performance. To test the performance of individual features, we experimented with all features but with one feature removed at once. The best feature is the translation probabilities because when it is removed, we observe the largest decrease in NCE. The translation source, the source sentence-related features and the language model features provide some marginally useful information.

5. CLIR EXPERIMENTS

In order to validate our confidence model for CLIR, we use English queries to retrieve French documents. We use two document collections: one from TREC6 and another from CLEF (SDA). We use 4 query sets: 3 from TREC (TREC 6, 7, 8) and one from CLEF2000. The query terms are translated with three sources: two STMs trained respectively from the Hansard and the Web corpus and one BD (http://www.freedict.com/). The resulting translations are then submitted to the information retrieval process. We tested with two ways to assign weights to translation candidates: linear combination and confidence estimation. In linear combination, each resource is assigned a coefficient denoting our confidence in it. In confidence estimation, we use confidence estimates as weights for translations instead of original probabilities. According to these confidence measures, we select the three translations with the best confidences for each query term. The following table shows the results:

Table1: Comparison of CLIR Performance between Linear Combination (LC) and Confidence Measures (CM)

Collection	TREC6	TREC7	TREC8	CLEF
MAP of LC	0.2692	0.2630	0.3605	0.3071
MAP of CM	0.2988	0.2699	0.3761	0.3230
Improvement rate of CM compared to LC	10.99	2.62	4.32	5.17

In terms of Mean Average Precision (MAP), we see clearly that the results using confidence estimation are higher than those obtained with the linear combination on all the collections. This improvement in CLIR performance is attributed to the ability of confidence measure to re-weight each translation candidate. The final set of translation words (and their probabilities) are more reasonable than in linear combination. Linear combination assumes that all the suggested candidates are correct and it simply groups them together. On the contrary, the confidence model does not blindly trust all the translations. It tests their validity on new validation data and according to new features. Thus, the translation candidates are rescored and filtered according to a more reliable weight. This strongly impacts the effectiveness of CLIR. Therefore, confidence measure provides a promising mechanism to select the most appropriate translations of a query.

6. CONCLUSION

Most previous studies in CLIR used a linear combination. This method is unable to combine correctly non homogeneous resources such as BDs and STMs. In this study, we examined the possibility of using a confidence estimation technique for the query translation task. This method reconsiders each translation candidate proposed by different resources with respect to additional features. It is able to re-weight the translation candidate more radically than in linear combination. This approach can be further improved on several aspects. For example, we can optimize this technique by identifying other informative features. Other techniques for computing confidence estimates can also be used in order to improve the performance of CLIR.

7. REFERENCES

[1] Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. Confidence estimation for machine translation. Technical Report, CLSP/JHU 2003 Summer Workshop, Baltimore, 2003.

[2] Chen, J., and Nie, J. Y. Parallel Web text mining for cross-language. RIAO, Paris, pp. 62-77, 2002.

[3] Nie, J. N., Simard, M., and Foster, G. Multilingual information retrieval based on parallel texts from the Web. In LNCS 2069, C. Peters editor, CLEF2000, pages 188-201, Lisbon 2000.

[4] Oard, D. W., and Dorr, B. J. Survey of multilingual text retrieval. UMIACS-TR-96-19 CS-TR-3815, 1996.

[5] Xu, J., and Weischedel, R. Empirical studies on the impact of lexical resources on CLIR performance. Information processing & management, 41(3), 475-487, 2005.