WeMiT: Web-Mining for Translation

Mathieu Roche [1] and **Oana Mihaela Garbasevschi** [2]

Abstract. The quality of machine translation is often dependent on the quality of lexical transfer from a source language to a target language. In this work we present an automatic method to translate specialized terms. The proposed approach is based on two steps: (1) extraction of candidates for translation into web pages, (2) identification of the most relevant candidates by using web-mining techniques.

1 INTRODUCTION

For translation task, the lexical transfer from one language to another one is crucial. However, current tools can be inefficient. For instance the French term *fouille du web* is often translated with *searching the web* (e.g. using *Google Translate* as the example of the Figure 1[3]). Of course this translation is irrelevant. Actually a lot of available tools have problems to translate phrases from specialized domains [3].

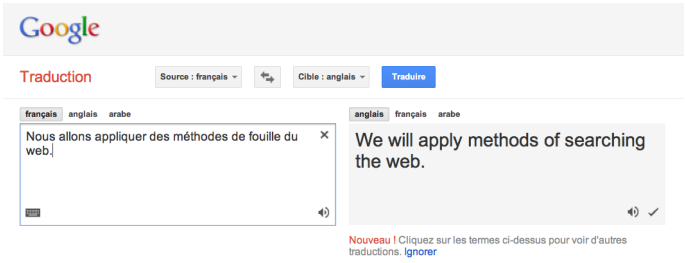

Figure 1. Example of google translation.

Generally the multilinguism lexical acquisition tasks are based on the use of alignments [6] or comparable corpora [8]. Other approaches use Wikipedia articles available in different languages [4]. In addition, the statistics of the web can be used to validate possible translations [8]. We will also use the web resources in order to extract translation but also to validate them. From this last task our method is closer to [9].

Our approach, called WeMiT (Web-Mining for Translation), allows to provide a relevant translation for a given term. WeMiT is based on the principle of PMI-IR algorithm (Pointwise Mutual Information and Information Retrieval) [9]. PMI-IR queries the Web using the AltaVista search engine in order to determine synonyms. In our approach, three major differences are identified. First, we apply different statistical measures to rank the elements. In addition, we use measures that research co-occurrences found in a context flexible or not. Finally, our approach is more global because it takes into account a preliminary step to extract candidates into Web pages. This point is developed in the next section. Ranking functions for translation are described in Section 2.2. Experiments on real data are developed in Section 3. Finally, Section 4 details the future work.

[1] LIRMM, CNRS, Univ. Montpellier 2, France, email: mroche@lirmm.fr
[2] Univ. Montpellier 2, France
[3] Test date: January 23, 2012

2 THE WeMiT APPROACH

2.1 Extraction of candidates for translation

In order to extract translation candidates from Web pages, we deal with the first 100 pages returned by a search engine (by specifying a language) with a query using the expression to translate *exp*.

To extract the candidates (*cand*), we adopt the following process.[4] For each page, we identify the parts where the expression is. We seek a first pair of parentheses in the text to extract its contents (e.g. *La fouille du Web (Web Mining, WM)*). In fact we assume that this marker (i.e. parenthese) is often adapted in order to find translation candidates. This type of method is also used for other tasks such as extraction of acronym/definition in texts [7]. After checking if this content is written in English, a cleaning process is applied (i.e. removing noise and linguistic markers as *called*, *too*, and so forth). So we have a list of candidates for translation according to the expression *exp*. For example with the expression to translate *fouille du web* (in French), we have obtained the candidates *open mango*, *web data*, *mailing*, *web mining*, *web mining wm*. The next section presents our approach to rank them.

2.2 Ranking of candidates

In order to rank candidates, we use four types of statistical measures that calculate the dependance between *exp* and *cand*.

Several measures can be applied in a web context developed in this work [1, 2, 7]. We select the more popular measures only based on the number of pages returned with *exp*, *cand*, and their co-occurrences:

- Frequency (FR): $nb(exp, cand)$

[4] In our experiments *exp* is a French expression and *cand* is an English candidate.

- Mutual Information (MI): $\frac{nb(exp,cand)}{nb(exp)\times nb(cand)}$
- Cubic Mutual Information (MI3): $\frac{nb(exp,cand)^3}{nb(exp)\times nb(cand)}$
- Dice Measure (DM): $\frac{2\times nb(exp,cand)}{nb(exp)+nb(cand)}$

Note that we use two types of co-occurences to calculate *nb*: (1) a strict co-occurrence to calculate the number of web pages containing the terms *exp* and *cand* one beside the other[5], (2) a flexible co-occurrence to calculate the number of times where *exp* and *cand* are in same pages.

Using the example of the previous section, we obtain the following values with Dice measure (with flexible *nb*) applied to the term *fouille du web* to translate. In particular, two candidates are possible: *web mining* and *web data*.
The following example shows that the translation *web mining* is more adapted:

- $DM(\text{fouille du web, web mining}) = \frac{2\times nb(\text{fouille du web,web mining})}{nb(\text{fouille du web})\times nb(\text{web mining})} = \frac{2\times 520}{9890+469000} = 0.0022$
- $DM(\text{fouille du web, web data}) = \frac{2\times nb(\text{fouille du web,web dada})}{nb(\text{fouille du web})\times nb(\text{web data})} = \frac{2\times 166}{9890+3180000} = 0.0001$

A graphical user interface has been developed to find a new translation online and/or to enrich a dictionary with terms (see Figure 2).

Figure 2. WeMiT Software.

3 EXPERIMENTS

In order to evaluate our methods applied in a French/English translation context, this section provides an evaluation of 358 couples (*exp*, *cand*). We have used a set of terms based on specialized documentations from Computer Science domain. For these experiments we have performed more than 1,500 queries with Google search engine.
To assess the measure quality the sum of the ranks of relevant translations is calculated.[6] The minimization of this sum is equivalent to maximize the Area Under the ROC Curve [5]. This principle is often used in data-mining field to assess the quality of ranking functions.

Table 1 presents the average of ranking sum obtained. The results show that strict dependencies are more efficient. Moreover, these results show that Dice measure (DM) has a good behavior with both types of dependencies (strict and flexible). With these paramaters (strict dependence + DM) based on 358 couples, 83% of the first translations returned with our system are relevant. With this same data set, the result given with *Google Translate* is 67%.

Strict dependence				Flexible dependence			
MI	MI3	DM	FR	MI	MI3	DM	FR
2.42	2.42	**2.28**	**2.28**	6.71	6.85	**6.14**	13.14

Table 1. Evaluation of measures with 127 couples.

4 CONCLUSION AND FUTURE WORK

In this paper, we have presented the WeMiT method which (1) extracts translation candidates from web pages, (2) ranks these translations with web-mining techniques.

Our system is based on an unsupervised approach. Supervised techniques could improve results. But in this case it is necessary to label manually a learning set with a high human cost. So in order to combine these different constraints, the use of active learning approaches could be adapted.

In our future work, we plan to combine strict and flexible dependencies with our web-mining approaches. Indeed, candidates can return no result with the strict dependence which is very restrictive. Thus, we propose to introduce a measure that ranks candidates by using strict dependencies, and when we obtain a score at zero, a flexible dependence will be applied. This principle takes into account the quality of results returned with strict dependencies and high coverage obtained with flexible dependencies. Finally we plan to propose other kinds of combinaisons too.

REFERENCES

[1] D. Bollegala, Y. Matsuo, and M. Ishizuka, 'Measuring semantic similarity between words using web search engines', in *Proc. of WWW*, pp. 757–766, (2007).
[2] R. Cilibrasi and P. M. B. Vitanyi, 'The google similarity distance', *IEEE Transactions on Knowledge and Data Engineering*, **19**(3), 370–383, (2007).
[3] V. Claveau, 'Automatic translation of biomedical terms by supervised machine learning', in *Proc. of LREC*, (2008).
[4] M. Erdmann, K. Nakayama, T. Hara, and S. Nishio, 'Extraction of bilingual terminology from a multilingual web-based encyclopedia', *Journal of Information Processing*, **16**, 68–79, (2008).
[5] C. Ferri, P. Flach, and J. Hernandez-Orallo, 'Learning decision trees using the area under the ROC curve', in *Proc. of ICML*, pp. 139–146, (2002).
[6] F.J. Och and H. Ney, 'The alignment template approach to statistical machine translation', *Computational Linguistics*, **30**(4), 417–449, (2004).
[7] M. Roche and V. Prince, 'Managing the acronym/expansion identification process for text-mining applications', *Int. J. Software and Informatics*, **2**(2), 163–179, (2008).
[8] F. Sadat, M. Yoshikawa, and S. Uemura, 'Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval', in *Proc. of ACL*, pp. 141–144, (2003).
[9] P.D. Turney, 'Mining the Web for synonyms: PMI–IR versus LSA on TOEFL', in *Proc. of ECML*, pp. 491–502, (2001).

[5] Exact search by the use of quotation marks (") in our queries.
[6] Actually several possible translations can be relevant.