Evaluation of terminologies acquired from comparable corpora : an application perspective

Estelle Delpech

Evaluation of terminologies acquired from comparable corpora : an application perspective

Estelle Delpech
Université de Nantes - LINA FRE UMRS 6249
2 rue de la Houssinière BP 92208, 44322 Nantes Cedex 3, France
estelle.delpech(at)univ-nantes(dot)fr
Lingua et Machina
c/o Inria Rocquencourt, BP 105 Le Chesnay Cedex 78153, France
ed(at)lingua-et-machina(dot)com

Abstract

This paper describes a protocol for the evaluation of bilingual terminologies acquired from comparable corpora. The aim of the protocol is to assess the terminologies'added-value in a task of specialized translation. The protocol consists in having specialized texts translated in various situations: without any specialized resource, with an domain-related bilingual terminology or using Internet. By comparing the quality of the segments translated using these various resources, we are able to assess the impact of our bilingual terminologies on the quality of the translation.

1 Introduction

Evaluation plays an important role in NLP developments: it assesses the quality of tools, brings out the progress made between two developments, spots the limitations and highlights possible lines of research. Regarding the evaluation of terminologies, (Nazarenko et al., 2009) show that terminologies are complex objects and that their evaluation can be quite arduous. These authors distinguish between three evaluation modes:

evaluation through reference : the terminology is compared to a standard reference, the evaluation metric indicates the adequacy between the assessed terminology and the reference terminology.

evaluation through interaction : the evaluation aims at measuring the cost of the transformation of the raw terminology as outputted by the system into the final, validated, ready-to-use terminology.

evaluation through application : the evaluation's purpose is to compare the performance of a given application with and without the terminology, the metric indicates the added-value of the terminology and depends on the application.

(Nazarenko et al., 2009) propose protocols and metrics for the first two evaluation modes and focus on monolingual terminologies only. The aim of this paper is to propose a protocol for the evaluation through application of bilingual terminologies acquired from comparable corpora. The considered application is human specialized translation.

Terminologies acquired from comparable corpora are usually assessed using an evaluation through reference protocol (Fung, 1997; Sadat et al., 2003; Koehn and Knight, 2002). Algorithms which extract bilingual terminologies from comparable corpora output a list of 1-to-n alignments: each source term is aligned with the *n* best candidate translations, most of the time the *Top20* candidate translations. The output of the algorithm is compared to a reference lexicon and the evaluation metric is a precision score computed on the *Top1*, *Top10* or *Top20* candidates. For example, a 50% precision on the *Top20* candidates indicates that the correct translation is found among the first top 20 candidates for 50% of the source terms.

Although evaluation through reference is useful to monitor the effect of changes in the alignment algorithm and to compare the alignment techniques, we believe it is important to demonstrate the impact and the usefulness of terminologies and lexicons acquired from comparable corpora in real-life applications. (Renders et al., 2003) showed the influence of such lexicons on cross-lingual information retrieval. We would like to determine the added-value of these bilingual terminologies when they are used in a task of human specialized translation.

Section 2 explains how terms are aligned in comparable corpora and examines the issue of translation quality assessment. The evaluation protocol is defined in section 3. The experimentation and results are described in section 4. Perspectives and future work are discussed in section 5.

2 Background

In this section, we describe the algorithm used for term alignment (section 2.1) and give a brief state-of-the-art survey in translation quality assessment (section 2.2).

2.1 Term alignment from comparable corpora

Comparable corpora are sets of texts written in two languages which are not translations of each other but which share a substantial part of their vocabulary, mainly because they are topic-related. The major advantage of comparable corpora is that it is much more available than parallel corpora and enables the processing of unprecedented language pairs. It is also often argued that the target language texts found in comparable corpora contain more spontaneous / natural terms and expressions than in parallel corpora because the target texts are not translations and they have not been influenced by the language of the source text.

Term alignment from comparable corpora was initiated by the work of (Fung, 1997) and (Rapp, 1995). The alignment algorithm is based on distributional linguistics and considers that two terms are probable translations if they occurr in similar contexts. The context of a term T is represented by a vector indicating the number of times T cooccurs with each word within a given contextual window (for instance: three words on the left of T and three words on the right of T). The cooccurrence frequencies are normalized using the log-likelihood ratio (Dunning, 1993). Words in the source context vectors are translated into the target language using a bilingual seed lexicon. Then, the source and target vectors are compared using a similiarity measure such as the Cosine similarity measure. The most similar the vectors, the most likely the target and source terms are translations of each other. (Morin and Daille, 2009) report that the correct translation is to be found among the top 20 best candidates for 42% to 80% of the source terms depending on corpus size, on the complexity of the terms and whether the alignment is made using specialized corpora or general language corpora. As a consequence, the output lexicon is ambiguous and sometimes, the correct translation does not appear among the candidates.

2.2 Translation quality assessement (TQA)

Because we want to compare the quality of translations made by humans with and without bilingual terminologies, we need to find a way to assess human translation quality. If Machine Translation (MT) enjoys well-defined and rather consensual metrics to evaluate its quality, evaluation of human translation poses a real challenge. These two domains use different protocols for the assessment of translations. On the one hand, MT evaluation focuses on comparing the output of different MT systems. This evaluation is done in reference to one or several human translations. On the other hand, translation studies seek to assess the quality of a human translation on its own, without any reference to a standard translation. In fact, the only reference is the judge himself/herself.

2.2.1 TQA and machine translation

There are two ways of assessing machine translations. One is called *objective* or *automatic evaluation*. The other is called *subjective* or *human evaluation*.

In objective evaluation, translations are evaluated through a measure which is automatically computable and which has the advantadge of being reproducible. Examples of such measure are: BLEU (Papineni et al., 2002) which is based on the count of common n-grams between the assessed translation and reference translation(s) and METEOR (Banerjee and Lavie, 2005) which is similar to BLEU but leaves room for variation by including morphological variants and synonyms in the n-gram comparison. These metrics are in turn meta-evaluated by computing their correlation with human jugements. Though handy for the evaluation of MT systems on a daily basis, these metrics were not used in the shared translation tasks of the ACL Workshop on Statistical Machine Translation where they are perceived as imperfect substitutes of human assessment, see (Callison-Burch et al., 2009) and (Callison-Burch et al., 2010) for example.

MT evaluation campaigns led to the development of a series of protocols for what is called *subjective* or *human* evaluation of MT. Two evaluation

protocols stand out :

judgement task The judge grades each translation independently. The grading scale can be quite complex, like 5 points scales over two criteria (fluency, adequacy) or simple binary judgements (correct/incorrect).

ranking task The judge ranks several translation of the same source segment from worse to best, each translation being produced by a different system.

These protocols are meta-evaluated using inter- and intra- annotator agreement measures which gives some indication on the coherence of the judgements. Experiments by (Callison-Burch et al., 2007) show that annotation tasks which involve complex sets of categories (e.g. 5 points scales over adequacy and fluency vs. binary judgements) and larger segments (sentences vs. phrases) tend to be more time-consuming and to result in lower agreement. The ranking task is considered easier and less time-consuming than the judgement task. It also yields a higher annotator agreement.

2.2.2 TQA and translation studies

In translation studies, TQA is mainly used by the translation industry as a way to monitor the quality of its products. (Secară, 2005) gives an overview of various translation grids. Although there is no consensus, all grids follow more or less the same methodology. Translation errors are categorized (e.g. spelling, grammar, terminology) and each error type is assigned a certain cost which is proportional to its gravity. A passage of a given length is randomly selected from the translation under assessment. Errors are marked and the cost points add up. If the sum of the points exceeds some threshold, the translation is deemed unacceptable.

These grids are criticised by theoretical approaches to TQA - see (Williams, 2001) for instance - because they stick to the lexical and syntactic levels and do not take into account higher linguistic levels like discursive or argumental structures. They are also monolithic and supposed applicable to any kind of text whereas authors like (Reiss, 1971) have argued that the evaluation criteria and their weight should be adapted to the text's function.

3 Protocol

The evaluation protocol is based on the ranking and judgement tasks used in MT subjective evaluation. These tasks were chosen because of their relative simplicity (compared to traditional 5-points scales) which also results in more reliable judgements as shown by (Callison-Burch et al., 2007). Automatic evaluation metrics were discarded because their only advantage - reproducibility - is of no use in this kind of evaluation: the protocol includes a subtask which is not reproducible (the translation) which makes the overall evaluation non-reproducible anyway. Evaluation grids were also discarded because they are too complex to put in practice, difficultly available and scarcely documented.

The evaluation's protocol is as follows:

1. Translators translate specialized texts in three different situations which we call "situations of translation". These situations of translation share a common base of identical generic resources (two monolingual and one bilingual dictionnaries). Translation are made from second-language to native language:

   situation 0 : translate with *generic resources* only

   situation 1 : translate with generic resources + a *bilingual terminology* extracted from comparable specialized corpora.

   situation 2 : translate with generic resources + full access to *Internet* where the translator can find all sorts of translation aids

   Situation 0 acts as a baseline where the translator has no specialized resource. The terminology used in situation 1 is the terminology under assessment. In situation 2, the Web is considered as some sort of "super-" or "meta-" specialized resource, because the translator will have access to all the specialized lexicons and termbases that are available online.

2. Once the translations are done, translators note down the time they spent in translation as well as the terms or expressions that they found problematic to translate and which drove them to use a linguistic resource. They also note down which resources they used to make the translation.

3. For each problematic term, judges rank the translations produced in the different situations[1] (ranking task). They also judge each translation separately using three categories: exact, acceptable or wrong (judgement task).

4. The added-valued of the bilingual terminology (situation 1) is measured by the comparing the quality of the translations produced in situation 1 with the quality of the translations produced in situations 0 and 2.

We decided to restrict the evaluation to the problematic terms rather than evaluating the quality of the whole translation because it appears from works in translation studies (Williams, 2001; Reiss, 1971) that the overall quality of the translation of a text emerges from the complex interaction of various parameters (register, syntax, argumental structure, spelling, etc.) most of which terminologies have no influence upon. By focusing on the problematic terms and expressions, we isolate the part of the translation that terminologies are meant to improve. As a side effect, evaluating small segments also saves time and yields more reliable judgements as demonstrated by (Callison-Burch et al., 2007).

The judgement task is based on three categories presented in table 1. These categories were chosen in accordance with (Reiss, 1971) who states that the translation of "content-focused texts" (e.g. scientific and technical texts, manual for use...) should favor the transfer of the source text's meaning over the transfer of the source text's form. An *acceptable* translation is a translation which conveys the meaning of the source term. An *exact* translation is a translation which makes use of the expected, standard target term. In a way, the "meaning transfer" and "accurate form" criteria parallel the more classical "adequacy" and "fluency" criteria found in MT campaigns.

	meaning transfer	accurate form
exact	✔	✔
acceptable	✔	
wrong		

Table 1: Translation quality criteria

In order to leverage differences in the quality of the translations which would arise from the translator's expertise rather than from the quality of the language resource, each situation of translation is evaluated on the basis of texts translated by several translators. In turn, one has to be cautious that translators do not translate texts from the same domain in different situations of translation. Indeed, if a translator translates a text from domain A in situation 1, he/she must not translate a text from domain A in situation 2: there is a risk that the translator re-uses some terms'translations he/she has learnt in the previous situation.

A critical point when judging the translation of technical texts is that the judges often lack domain expertise and that domain experts are rarely available. One can get round this trouble by choosing specialized texts which already have an existing translation, like research paper abstracts for example. Judges can also get help from general terminological databases such as *Termium* [2].

The consistency of the judgements can be improved by first running a blank evaluation on a small set of data and then discussing the disagreements with the judges (Blanchon and Boitet, 2007). In any case, it is necessary to provide the judges with clear instructions and examples of annotations on debatable cases.

4 Experimentation

This section describes the experimental framework (section 4.1) and the result of the evaluation (section 4.2).

4.1 Experimental framework

4.1.1 Data

Two bilingual terminologies were built for the evaluation. One was acquired from comparable corpora on BREAST CANCER and the other from comparable corpora on WATER SCIENCE . The WATER SCIENCE corpus is quite large (2M words per language) and its topic is coarse-grained. Texts are research papers from the journals *Sciences de l'eau*[3] and *Water Science and Technology*[4]. Conversely, the BREAST CANCER corpus is small (400k words per language) with a fine-grained topic. Texts come from various research papers of the publications portal *Elsevier* [5]. The texts to be translated belong to the same domains. They are divided into scientific texts and popular science

[1] Ties are allowed.

[2] http://www.termiumplus.gc.ca/

[3] http://www.rse.inrs.ca/

[4] http://www.iwaponline.com/wst/

[5] http://www.elsevier.com/

texts. The scientic texts are 2 × 3 research papers abstracts taken from *Elsevier* and the water science journals. The popular science texts are 2 × 1 webpages taken from bilingual websites on breast cancer[6] and water treatment[7]).

	BREAST CANCER	WATER SCIENCE
scientific	508	499
pop. science	613	425

Table 2: Size of texts to be translated (number of words)

4.1.2 Data processing

The algorithm described in (Fung, 1997) was applied to the terms and to every open-class word occurring more than 5 times in the corpus. Extra knowledge was automatically added to the terms and open-class words in order to help the translators: part-of-speech, frequency, collocations[8], variants[9], related terms[10], definitions[11], concordances. Translators could browse the terminology via a dedicated interface designed for terminologies acquired from comparable corpora (Delpech and Daille, 2010).

4.1.3 People involved

Due to the lack of human resources to perform the evaluation, there was some collisions in the roles of translator/judge and translator/organizer. Three persons were involved in the test of the protocol. The author of the paper, who is not a trained translator, translated the texts in the baseline situtation (general resources only) and organized the evaluation. Two trained translators translated the texts in situation 1 (terminology) and 2 (Internet) and also judged the translations. The translations were anonymized and randomly shuffled so that the judges would not know the origin of the translations.

Texts, domains and situations were distributed as follows :

[6]http://www.cbcf.org/

[7]http://www.lenntech.com/

[8]most remarkable cooccurrents, the association measure is the log-likelihood ratio (Dunning, 1993)

[9]phrases which have not been identified as terms by the term extractor but have words in common with the entry term

[10]terms which have words in common with the entry term

[11]either the Wikipedia or Wiktionary article if avalaible or a sentence extracted from the corpus and containg a very simple pattern like "A $TERM is a..."

	BREAST CANCER	WATER SCIENCE
untrained translator	sit. 0	sit. 0
trained translator 1	sit. 1	sit. 2
trained translator 2	sit. 2	sit. 1

4.2 Results

4.2.1 Translators' feedback

It was difficult for translators to adapt to the ambiguity of the alignments. Although the aim and the context of the evaluation had been explained to them, they still expected the correct translation to appear "on click", just like it happens with the traditional languages resources they are accustomed to. Another obstacle was the coverage of the terminology, especially for the WATER SCIENCE domain whose topic was not refined enough. Table 3 shows the percentage of words of the texts to be translated which also appear in the terminology. Clearly, fined-grained corpora should be favored over large corpora.

	BREAST CANCER	WATER SCIENCE
EN texts	94%	14%
FR texts	67%	78%

Table 3: Terminology coverage of the vocabulary of the texts to be translated (EN) and their reference translation (FR)

4.2.2 Problematic terms

Problematic terms are terms or expressions that a translator found difficult to translate. Problematic terms retained for the evaluation are terms which were tagged problematic by at least 2 translators. We collected 148 problematic terms (26 tagged by 2 translators and 122 tagged by 3 translators). Table 4 shows the repartition of problematic terms among domains and types of corpora.

	BREAST CANCER	WATER SCIENCE
pop. science	34	10
specialized	43	51
total	87	61

Table 4: Problematic terms used for evaluation

4.2.3 Time

The texts to be translated amounted to 2,147 words. Translators were quicker in situation 0

which is normal because they had less resources to browse (7.15 words/sec. on average). There is no significant time difference between situation 1 and situation 2 (11.18 and 11.6 words/sec. respectively).

4.2.4 Agreement between judges

Agreement was computed using the Kappa coefficient (Carletta, 1996) which takes into account the observed agreement $P(A)$ and the agreement which would have occurred by chance $P(E)$.

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Agreement was better for the ranking task: 0.65 (substantial) than for the judgement task: 0.36 (fair) which is consistent with the findings of (Callison-Burch et al., 2007). Agreement was better for the popular science texts: 0.57 (moderate) than for the scientific texts: 0.48 (moderate).

4.3 Judgement task

Table 5 gives quality judgements for the translations of the BREAST CANCER texts. The proportion of translations judged *wrong* is almost equivalent for all situations. Translations produced in situation 1 (with the terminology) are more often judged *exact* than the translations produced in situation 0 (only generic resources). Translations produced in situation 2 (with Internet) are the most accurate ones.

	sit.0	sit.1	sit.2
exact	38%	43%	47%
acceptable	42%	38%	35%
wrong	20%	19%	18%

Table 5: Translations' quality - BREAST CANCER domain

Table 6 gives quality judgements for the translations of the WATER SCIENCE texts. One can see that translations produced in situation 1 are of lesser quality than those produced in situation 0. This is unexpected because situation 1 and situation 2 share a common base of generic resources. Translations produced in situation 1 should be at least as good as translations produced in situation 0.

The fact is that the translators used the languages resources in different manners depending on the situation in which they performed the translations. Thanks to the data collected during the translation phase, we are able to tell, for each term translation if it was produced using the generic resource or the specialized resource (terminology/Internet) or relying on intuition (not exclusive). Table 7 shows that the translators who had access to a specialized resource scarcely used the generic resource. It might be because they felt the generic resource was useless to get the translation of technical terms and they prefered to use directly the specialised resource. But as the WATER SCIENCE terminology covers only a small part of vocabulary of the texts to translate, it was barely advantageous. A systematic exploitation of the generic resource in situation 1 would have led to translations at least as good as those produced in situation 0.

	sit.0	sit.1	sit.2
exact	59%	56%	77%
acceptable	23%	23%	16%
wrong	18%	21%	7%

Table 6: Translations' quality - WATER SCIENCE domain

	sit.0	sit.1	sit.2
gen. ress.	43%	14%	3%
spec. ress.	-	25%	56%
intutition	79%	77%	44%

Table 7: Exploitation of the language resources depending on the situation of translation

4.4 Ranking task

The ranking task results are similar to those of the judgement task. When different translations of the same terms are compared, those produced in situation 2 are always better, whatever the domain. Those produced in situation 1 are better than the ones produced in situation 0 only for the BREAST CANCER domain, probably because of divergences in the exploitation of the language resources as explained above.

5 Discussion and future work

We have described a protocol which assesses the added-value of terminologies acquired from comparable corpora when used for specialized human translation. This protocol consists in comparing

	sit.0	sit.2
sit.1 better than	28%	26%
sit.1 as good as	47%	42%
sit.1 worse than	26%	32%

Table 8: Translations' ranking - BREAST CANCER domain

	sit.0	sit.2
sit.1 better than	18%	16%
sit.1 as good as	49%	41%
sit.1 worse than	33%	43%

Table 9: Translations' ranking - WATER SCIENCE domain

several situations of translation in which the translators have access to diverse language resources: only generic resources, generic resources and the evaluated terminology, generic resources and full access to Internet. The added-value of the terminology is supposed to be evidenced by the difference in the quality of the translations produced in the three situations. We have described in section 4 a first trial of the protocol. This first trial showed that some hitches in our procedure prevent us from clearly demonstrating the added-value of terminologies acquired from comparable corpora : we had contradictory results for the BREAST CANCER and WATER SCIENCE domains. Nonetheless, this first experimentation, although carried out with a small set of data and participants, allowed us to test the feasibility of the protocol and pinpointed problems which must be solved before launching a more thorough evaluation :

- The observed added-value of the terminology highly depends on its coverage of the texts used to evaluate it. Any measure of this added-value should also mention the adequacy between the assessed terminology and the texts to be translated, otherwise it is not interpretable. We determined this adequacy in a simple manner, by computing the proportion of words in the texts to be translated that also occurr in the terminology. This leaves some room for improvement. The comparability of the corpora used form terminology extraction and alignment must also be taken into account. For this, we are planning to use the comparability measure developped by (Bo and Gaussier, 2010).
- The joint use of several language resources seems to bias the results as the translators' behaviour changes in function of the resources he/she has as his/her disposal. It is better to have only one resource per situation of translation, for instance:
  - situation 0: no resources,
  - situation 1: assessed terminology only,
  - situation 2: Internet only.
- Translators should be prepared to translate in a situation which is unusual to them. Ideally, one shoud run at first a blank translation task so as to discuss it with the translators and help them apprehend these new situations and resources.

The next step is to scale-up the protocol. We will renew the experiment on a much larger scale (a whole class of students translators) and include all the improvements listed above.

Finally, even if it was not the goal of this work, this first evaluation gives rise to some lines of research to improve the usefulness of terminologies acquired from comparable corpora. First, we have seen that the acquisition corpus should be collected in function of the texts that are to be translated and that the topic should be fine-grained. Second, it is clear that the Internet is a huge repository of linguistic resources and translations. A nice development would be to add a new functionality to the terminology software which, when the queried term is not present in the database, would either automatically generate a translation and filter it on the Internet or search it in pre-selected online resources. However, the worth of Internet as a linguistic resource should not be overestimated. In most professional translations, translators have to translate texts whose vocabulary can not be found on the Internet. It is especially the case with corporate translations : companies use their own terminologies, which can only be found in the texts produced by the company itself. Thus, we can not expect to rely on Internet as a unique source of translations and still need to improve the term alignment program. For this, we are planning to use translation techniques relying on the compositionality of terms (Morin and Daille, 2009) in addition to the distribution-based approaches (Fung, 1997) presented in section 2.1 and which we used for this evaluation.

Acknowledgments

This work was funded by the company Lingua et Machina and the French National Research Agency (funding no. ANR-08-CORD-009). I would like to thank Clémence de Baudus and Mathieu Delage for their participation in the evaluation.

References

S. Banerjee and A. Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Ann Arbor, Michigan.

H. Blanchon and C. Boitet. 2007. Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche. *Traitement Automatique des Langues*, 48(1):33–65.

L. Bo and E. Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *23ème International Conference on Computational Linguistics*, pages 23–27, Beijing, Chine.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the 2nd workshop on Statistical Machine Translation*, page 136–158, Prague, Czech Republic.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 17–53, Uppsala, Sweden.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

E. Delpech and B. Daille. 2010. Dealing with lexicon acquired from comparable corpora : validation and exchange. In *Proceedings of the 2010 Terminology and Knowledge Engineering Conference (TKE 2010)*, pages 211–223, Dublin, Ireland.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

P. Fung. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Unsupervised Lexical Acquisition: Proceedings of the Workshop pf the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9–16, Philadelphia, Pennsylvania. Association for Computational Linguistics.

E. Morin and B. Daille. 2009. Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation (LRE)*, volume 44 of *Multiword expression: hard going or plain sailing*, pages 79–95. P. Rayson, S. Piao, S. Sharoff, S. Evert, B. Villada Moirón, springer netherlands edition.

A. Nazarenko, H. Zargayouna, O. Hamon, and J. Van Puymbrouk. 2009. Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement Automatique des Langues*, 50(1):257–281.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

R. Rapp. 1995. Identifying word translations in Non-Parallel texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Boston, Massachussets, USA.

K. Reiss. 1971. *Translation criticism, the potentials and limitations : categories and criteria for translation quality assessment*. St. Jerome Pub., Manchester, GB.

M. Renders, H. Djean, and E. Gaussier. 2003. Assessing automatically extracted bilingual lexicons for CLIR in vertical domains: XRCE participation in the GIRT track of CLEF 2002. *Lecture Notes in Computer Science*, 2785/2003:363–371.

F. Sadat, M. Yoshikawa, and S. Uemura. 2003. Learning bilingual translations from comparable corpora to Cross-Language information retrieval: Hybrid statistics-based and linguistics-based approach. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, volume 11, pages 57–64, Sappro, Japan.

A. Secară. 2005. Translation evaluation - a state of the art survey. In *eCoLoRe / MeLLANGE Workshop*, pages 39–44, Leeds, UK.

M. Williams. 2001. The application of argumentation theory to translation quality assessment. *Meta : journal des traducteurs / Meta: Translator's Journal*, 46(2):326–344.