Bike: Bilingual Keyphrase Experiments

David Nadeau, Caroline Barrière and George Foster

Institute for Information Technology
National Research Council Canada
Ottawa, Ontario, Canada
{firstname.lastname}@nrc-cnrc.gc.ca

Abstract: This paper presents a novel strategy for translating lists of keyphrases. Typical keyphrase lists appear in scientific articles, information retrieval systems and web page meta-data. Our system combines a statistical translation model trained on a bilingual corpus of scientific papers with sense-focused look-up in a large bilingual terminological resource. For the latter, we developed a novel technique that benefits from viewing the keyphrase list as contextual help for sense disambiguation. The optimal combination of modules was discovered by a genetic algorithm. Our work applies to the French / English language pair.

Keywords: statistical machine translation, lexical resources, keyphrase list.

1 Introduction

As evidenced by recent comparative evaluations such as NIST05[1], most research in machine translation (MT) focuses on translating texts at the sentence level or above. In contrast, our work concerns sub-sentence-level translation; specifically, translating semantically-coherent lists of keyphrases. This task is related to query translation for cross-language information retrieval, and cross-language summarization (assuming a list of keyphrases constitutes the summary).

We built a bidirectional system for French and English, although the experiments reported here were conducted only on English to French translation. Our basic strategy is to combine a statistical translation model with a sense-focused look-up system based on a terminological resource which uses the list of keyphrases as context. Furthermore, to specifically target keyphrase translation, we applied heuristics that range from the validation of the correctness of keyphrase constructs to the handling of inflectional morphology.

The paper is organized as follows. We first survey related literature. Next, in section 3, we give a formal definition of our task. Section 4 presents experiments on each of the main modules in isolation. Section 5 describes the hybrid system we built to combine all modules. Section 6 concludes.

2 Related Work

Although a less active field than sentence-level MT, sub-sentence MT has nonetheless received a fair amount of attention, both for tasks like query translation in cross-language information retrieval, and term translation as a subtask of sentence-level translation.

Koehn and Knight [2003] address the problem of translating maximal noun and prepositional phrases as a subtask of translation. They train a maximum entropy model on a corpus of noun phrase translations, using features like the counts of noun, preposition, and determiner matches. Another approach [Moore, 2003] devoted to translating named entities involves adding capitalization and lexical information to classical translation models. Other examples of sub-sentence translation modules include [Kupiec, 1998] and [Collier *et al.*, 1998].

[1] www.nist.gov/speech/tests/mt

Fung and McKeown [1997] describe an algorithm for extracting term translations from parallel corpora (unaligned and possibly noisy), intended to serve as a translator's tool. They use anchor points and distances between multiple occurrences of the same term to induce translations. This approach takes us closer to the use of lexical resources, as it in fact uses statistical techniques to automatically generate a bilingual lexicon.

The use of existing (hand-built) lexical and terminological resources in sub-sentence translation is less common, due to the limited availability of resources, and to the problem of polysemy [Sadat *et al*. 2001]. In our work, we propose a novel technique for this problem, making use of the keyphrase list as a context for disambiguation. To our knowledge, this is the first use of the Grand Dictionaire Terminologique[2] (GDT) for automatic translation, although previous work has been done using the EuroWordNet thesaurus[3], the Babylon bilingual dictionary[4], and Termium [Nadeau et al. 2004] among other resources. Lexical and terminological resources can also be combined with other sources of information, as in [Cao and Li, 2002] where noun phrase translation is performed using a dictionary and the web. The dictionary gives word-to-word translation, while the web serves in finding the best multi-word arrangement.

Combination of multiple resources (typically statistical and symbolic techniques) has often proved useful in query translation for cross-language information retrieval. Hibash and Dorr [2002] start by generating all possible translations from a symbolic resource and then use a statistical model to select the best in context. This strategy is often used as a combination technique but many other approaches have been proposed, as can be appreciated from the proceedings of a workshop dedicated to the task [Klavans and Resnik, 1994].

In this work, we also propose a combination of resources aimed at translating keyphrases. We combine symbolic and statistical information using a genetic algorithm to optimize six parameters that describe the contribution of each module. Our work differs from previous work in the use of the genetic search, and in exploiting keyphrase list context for disambiguation.

3 Definition of the Task

The task consists in translating a list of English keyphrases into its French equivalent. The keyphrases are drawn from a collection of scientific papers taken from the Canada Institute for Scientific and Technical Information (CISTI). Papers are from ten journals in the following domains: biochemistry, botany, chemistry, civil engineering, environment, genomics, geotechnical, microbiology and pharmacology.

Each paper contains an abstract (A) and a list of keyphrases (K) both available in French (f) and English (e). Most papers are originally in English and the A_e and K_e are translated to A_f and K_f by CISTI editors. A minority of papers is originally in French and, reciprocally, A_f and K_f are translated to A_e and K_e by CISTI editors.

The collection is made up of 3,058 document tuples $\{A_f, A_e, K_f, K_e, T\}$. Note that the source text (T) is either French or English but never translated. The A_f average size is 258 words, and the A_e average size is 207 words. The average number of French keyphrases (K_f) is 11, and the average number of English keyphrases (K_e) is 9.

The metric we use in this paper is accuracy: the ratio of correctly translated keyphrases to the total number of keyphrases. For reproducibility, we consider an

[2] http://www.granddictionnaire.com
[3] http://www.illc.uva.nl/EuroWordNet/
[4] http://www.babylon.com/

English keyphrase to be correctly translated only if it is an exact match with the corresponding French reference (human) translation. The results we give are for the global corpus, but we also calculated the accuracy by journal in order to perform significance tests based on the mean per-journal accuracy.

We used 40% of the corpus for training (for statistical MT), 30% for testing as well as performing a genetic search for the combination of approaches. The final 30% of "held-out" was used at the very end of the experiments to report the results.

4 Experiments with Isolated Methods

In this section, we report the accuracy of various methods used in isolation: first a baseline experiment, followed by the statistical MT module and, finally, the terminological resource approach.

4.1 Baseline Approach

The baseline experiment consists simply in not translating the keyphrases. Since French and English share many words, this technique achieves a non-null accuracy, as seen in Table 1.

	Accuracy (%)
Baseline experiment	20.21

Table 1: Result for the baseline experiment.

The baseline idea is used as a default strategy throughout our experiments when no suggestions can be made by the systems.

4.2 Statistical Translation Model

The statistical translation model is a joint distribution over French, English phrase (ngram) pairs as used in phrase-based statistical MT [Koehn *et al.*, 2003]. The model is induced from a sentence-aligned parallel corpus in two main steps. First, perform a word alignment for each sentence pair using IBM2 models [Brown *et al*., 1993], merging the alignments from English-to-French and French-to-English directions with heuristics similar to those described by Koehn *et al.* Second, extract all phrase pairs (up to a maximum of 8 words per language) that are consistent with the word alignment, in the sense that there are no links external to the pair. All extracted pairs are added to the distribution, and probabilities are assigned using simple relative-frequency estimates.

To use this model for keyphrase translation, we calculate conditional French-given-English probabilities, and retain only the most probable translation for each English phrase. Translation is then simply a matter of finding a verbatim match for an input keyphrase (if one exists) and outputting the corresponding French translation

We first tested a statistical MT trained on an external corpus, the Canadian Hansard. We called it the "hansard" model. We then trained a translation model on the CISTI corpus, using sentence-aligned abstracts and keyphrase lists. We refer to this as the "cisti" model. In the latter case, two strategies were used: creating a single global model for all journals, and creating an individual model for each journal (since we know the domain at run time, we can select the appropriate individual model). Results are presented in Table 2:

	Accuracy (%)
hansard (with baseline)	10.47 (26.11)
cisti global (with baseline)	30.58 (39.54)
cisti individual (with baseline)	25.27 (36.50)

Table 2: Results for the MT models.

The poor performance of the cisti individual models appears to be due to the sparseness of the training material; the improvement achieved by the cisti global model means that there is a set of phrase translations that are relatively domain independent.

4.2.1 Enhancement: MT Correction

A shortcoming of the statistical MT approach is inherent to the use of phrase-based translation within sentence-level translation. At sub-sentence level, French nouns or noun compounds tend to be surrounded by extras determiners, punctuations or prepositions. Table 3 shows some examples:

English source	French equivalent (s)
fiber	fibre, de fibre
population sizes	des dimensions des populations

Table 3: Examples of determiners added to French nouns

The easy way to solve this problem is simply to remove prefixes (`"de la ", "le ", "la ", "les ", "l'", "du ", "de ", ","`) and suffixes (`" ," , " de", " du", " des"`) from French translation proposed by the MT system. This simple correction, as shown in Table 4, improved by more than 1% the best result shown in Table 2, a result that is statistically significant at the 95% level, according to the standard paired t-test. For the remaining experiments, this correction is always applied for statistical MT translations.

	Accuracy (%)
cisti global model + correction	41.26

Table 4: Results for the cisti MT model using the correction heuristic.

4.3 Terminological Resource

The terminological resource we use is the Grand Dictionaire Terminologique (GDT). In case of polysemous words, as a first naive approach, we choose the first record of GDT, corresponding to the most common sense. Extracted keyphrases from scientific journals are often terms[5] and therefore likely to be found in a terminological resource, as we can see from Table 6 (section 4.3.1). Some sample entries, shown in Table 5, illustrate the content of GDT:

English source	French equivalent (s)
field testing	essai pratique, test sur place
field theory	théorie des champs, théorie du champ, théorie de l'influence du milieu sur le comportement

Table 5: Sample entries in GDT.

[5] Terms are usually noun compounds with a particular meaning within a domain.

4.3.1 Morphology

A problem with the terminological resource is that an input keyphrase must exactly match an entry in order to yield an equivalent in the other language. To deal with the fact that the vast majority of entries are singular, we created a simple morphological analyzer to map plural English keyphrase inputs to singular forms, and the resulting translations back into plural French forms.

To transform English words to singular form, we handle endings[6]: `"xes"`, `"ches"`, `"sses"`, `"shes"`, `"[^aeiouy]ies"`, `"[^f]ves"`, `"[lr]ves"`, `"ses"`, `"men"` and `"s"`. For multi-word expressions, we simply applied this ending filter on the last word of the expression.

To transform back to French plurals, we handle endings: `"al"`, `"eau"`, `"s"` and `"x"`. Other endings are simply extended with an `"s"`. For multi-word expressions, we pluralize every single word, except for stop words like `"de"`, `"d"`, `"en"`, `"à"`, `"par"`, `"sur"`. Also, when those stop words are encountered, we do not pluralize subsequent words of the expression.

This basic morphology algorithm improves results for the GDT, as shown in Table 6. Such approach is not necessary with statistical MT since, by design, it already handles plural and singular forms.

	Accuracy (%)
GDT	35.66
GDT + morphology	38.30

Table 6: Results for the GDT with and without the morphology heuristic.

The morphology improvement is significant at the 99% level. In the rest of the paper, all experiments involving the GDT include the morphology module.

4.3.2 Domain Attribution

The domain attribution approach is the cornerstone of the use of the terminological resource. It is used to partially solve the problem of polysemy, when an English keyphrase has many possible records in the GDT.

A particularity of all terminological databases (GDT included) is that terms are classified by domain, such as biology, geology, transport, food industry, environment protection, zoology, tobacco. These domains are often organized hierarchically according to their specificity. For the present research, the GDT's top level domains (about 200) are used.

Finding the right translation for a keyphrase in the GDT means finding its appropriate entry given the context of use. This boils down to a word sense disambiguation problem, and since GDT categorizes terms by domains, we can reformulate it as a domain disambiguation problem.

The idea of domain attribution is to find a minimal set of domains covering every keyphrases. We have devised a Minimal Domain Set (MDS) algorithmic solution to find, among all the domains covered by the keyphrases, which set gives 1) the most likely and 2) the most coherent group of domains.

Coherence is estimated by domain similarity, measured as follow. For all pairs of domains, similarity between a pair is the number of records (word senses) they share.

[6] Some endings are in the regular expression format.

Likeliness is measured by coverage (a domain is more likely if it covers more keyphrases in the list). Here is the detailed algorithm.

1. Initialize the frequency *F(D)* of each domain to 0.
2. L := list of keyphrases.

3. For each keyphrase $K_{i=0}^{i=|L|}$:
3.1. Increment the frequency *F(D)* of each domain of K_i.
3.2. $|D_{Ki}|$:= number of domains of K_i.

4. L' := Sort L in ascending order of $|D_{Ki}|$.

5. Initialize the empty Minimal Domain Set (MDS).

6. Loop until MDS covers at least a domain of each keyphrase:

6.1. For each keyphrase $K_{i=0}^{i=|L|}$:
6.1.1. **Likeliness**: From the list of domains D_{Ki}, build a reduced list containing only the domains with the highest frequency *F(D)*.
6.1.2. **Coherence**: From this reduced list, select the domain which has the highest coherence with a member of MDS[7]. Add this domain to MDS.

Once MS is built, keyphrase sense disambiguation is performed by choosing the sense (record) in the GDT which corresponds to one domain in MDS. In the few cases where two records share the same domain, we arbitrarily take one, as we do not proceed to any further sense disambiguation beyond domain disambiguation[8]. Furthermore, if a record contains more than a single French equivalent (sometimes synonyms are given), the most common keyphrase equivalent is chosen (commonality being approximated by hit count on Google). Using this strategy, the GDT result improves from an accuracy of 38.30% to 39.22%. The improvement is significant at the 95% level.

	Accuracy (%)
GDT + Domain attribution	39.22

Table 7: Results for the GDT using the domain attribution.

At this point, the difference between the best MT module (Table 4) and the best GDT setup (Table 7) is not significant (below 95%). The absolute margin is ~2% but GDT dominates MT on half the journals and reciprocally on the other half.

5 Experiments with Hybrid Approaches

We examined two methods for combining the statistical MT approach and terminological resource approach.

[7] If MDS is empty, select a domain at random. In our experiment, this case never occurs since MDS is only empty at the first iteration and there is usually only one domain at step 6.1.1 and 6.1.2 because of the sorting performed at step 4.

[8] It is rather unlikely to have a term representing two notions (senses) within the same domain, but it does happen.

5.1 Module Cascade

We first test the idea of using the modules in cascade. That means translating with a module and using a second module should the first one have no output, and so on for each module. The best sequence of modules, for which result is shown in Table 8 is: the MT cisti global model followed by the GDT and the MT hansard model.

	Accuracy (%)
Cascading cisti global model, GDT and hansard model	43.83

Table 8: Results of cascading modules

5.2 Weighted Choice

To combine modules in a more sophisticated way, we used a weighting model involving six parameters. The score assigned to each hypothesis is the product of its base score (probability in the case of the MT module; 1 over number of translations in the case of the GDT) times six Boolean feature weights. Each weight is set to 1 if the corresponding feature does not apply, otherwise to a learned value. Features are the following (the number in brackets is optimal weight):

- [1.071] candidate proposed by MT cisti model;
- [0.227] candidate proposed by MT hansard model;
- [0.477] candidate proposed by GDT;
- [1.464] candidate proposed by more than one source;
- [0.853] candidate transformed to singular (see 4.3.1);
- [1.257] GDT candidate(s) in MDS (see 4.3.2);

To find optimum weight values, we used a standard genetic algorithm designed to run on a computer cluster. We executed the optimization task on a 20-node cluster, allowing us to test several parameter sets and run near-exhaustive experiments. The best result found, shown in Table 9, is an improvement over the module cascade statistically significant at the 99% level.

	Accuracy (%)
Best weighted model	47.16

Table 9: Best weighted model found by genetic search.

6 Conclusion

In this paper, we explore the task of translating a list of keyphrases. Unlike sentence-level MT, it requires some heuristics to adapt the classical statistical models. This task also benefits from the contribution of terminological databases such as the Grand Dictionnaire Terminologique. We show how to combine multiple translation modules and how to apply simple heuristics designed for the keyphrase list translation problem. We present a novel idea to use the list of keyphrases as a context for disambiguation of polysemic keyphrases. Our final system outperforms any module taken in isolation on the CISTI corpus.

Future work may require improvement on both main modules, first to further modify the statistical MT idea to have an intrinsic correction of prefixes and suffixes in the phrase table (see section 4.2.1), and second to augment the number of keyphrases for which the GDT could provide translations by looking into individual words of the often multi-word keyphrases.

7 Acknowledgement

We thank the *Office québécois de la langue française* for granting us access to the entire GDT database. We also acknowledge the collaboration of the Canada Institute for Scientific and Technical Information (CISTI) that supplied the corpus, and the financial support from the Official Languages Innovation Program of Canada.

8 References

[Brown *et al*., 1993] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993). The Mathematics of Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 1993.

[Cao and Li, 2002] Yunbo Cao and Hang Li (2002). Base Noun Phrase Translation Using Web Data and the EM Algorithm. Coling02.

[Collier *et al*., 1998] Nigel Collier, Hideki Hirakawa, and Akira Kumano (1998). Machine Translation vs. Dictionary Term Translation - a Comparison for English-Japanese News Article Alignment. ColingACL98.

[Fung and McKeown, 1997] P. Fung and K. McKeown (1997). A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups. Machine Translation

[Koehn and Knight, 2003] Philipp Koehn and Kevin Knight (2003). Feature-Rich Statistical Translation of Noun Phrases. ACL03.

[Koehn *et al*., 2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu (2003). Statistical Phrase-Based Translation. NAACL03.

[Kupiec, 1998] Julian Kupiec (1993). An Algorithm For Finding Noun Phrase Correspondences In Bilingual Corpora. ACL93.

[Moore, 2003] Robert C. Moore (2003). Learning Translations of Named-Entity Phrases from Parallel Corpora. EACL03.

[Sadat *et al*. 2001] Fatiha Sadat, Akira Maeda, Masatoshi Yoshikawa, and Shunsuke Uemura 2001. Query Expansion Techniques for the CLEF Bilingual Track. In Proceedings of the CLEF 2001 Workshop on Cross-language System Evaluation Campaign, pp. 99-104, Darmstadt, Germany.

[Hibash and Dorr, 2002] Nizar Hibash and Bonnie Dorr (2002). Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. AMT02.

[Klavans and Resnik, 1994] Judith Klavans and Philipp Resnik (1994). The Balancing Act: Combining Symbolic and Statistical Approaches to Language. ACL Workshop Proceedings.

[Nadeau et al. 2004] David Nadeau, Mario Jarmasz, Caroline Barrière, George Foster and Claude St-Jacques (2004). Using COTS Search Engines and Custom Query Strategies at CLEF. Cross-Language Evaluation Forum CLEF 2004.