RBMT as an alternative to SMT for under-resourced languages

Guillaume de Malézieux
INaLCO, Paris
guillaume2l2m@gmail.com

Amélie Bosc
INaLCO, Paris
amelie.bosc@gmail.com

Vincent Berment
INaLCO, Paris
LIG/GÉTALP, Grenoble
Vincent.Berment@imag.fr

Abstract

Despite SMT (Statistical Machine Translation) recently revolutionised MT for major language pairs, when addressing under-resourced and, to some extent, mildly-resourced languages, it still faces some difficulties such as the need of important quantities of parallel texts, the limited guaranty of the quality, etc. We thus speculate that RBMT (Rule Based Machine Translation) can fill the gap for these languages.

1 Introduction

In this paper, we present an ongoing work that aims at assessing the relevance of specific methods to reach "quick and quality" machine translation for under-resourced languages. These methods include working in parallel on several languages, reusing software and linguistic resources, relying on a pivot architecture, opening our linguistic sources and letting any group of users the possibility to "do it themselves". We also chose to adopt the old fashioned RBMT approach.

More concretely, we are applying Vauquois' methodology [Vauquois and Chappuy, 1985] to the development of analysers for Khmer, Lao, Thai and Hindi, which we plan to "connect" to existing and open source syntheses of French and English through three means: deep transfer, deep hybrid transfer and UNL pivot representation. In order to elaborate easy-to-understand guidelines for new comers, we chose to create a primer methodological step involving the small novel of Saint-Exupéry "The Little Prince", which has been translated into 270 languages and dialects. Doing so, the principles for developing dictionaries and grammars that follow Vauquois' methodology become much simpler to understand.

2 Tools and methodology

2.1 The Heloise RBMT framework

The RBMT framework we are using is called Heloise. It has been presented at COLING 2012 [Berment and Boitet, 2012]. Heloise is an online environment available to anyone wishing to design his or her own operational expert MT system, especially for under-resourced pairs of languages. It is upward-compatible with Ariane-G5's languages, so the open-source modules developed under this environment can be reused in any new system. For example, in order to add a new language X, an existing generation of French language can be taken as such for a new X-French system, limiting the effort to an analyser of language X and to a transfer from X to French. Figure 1 represents the usual phases involved in a development under Ariane-G5.

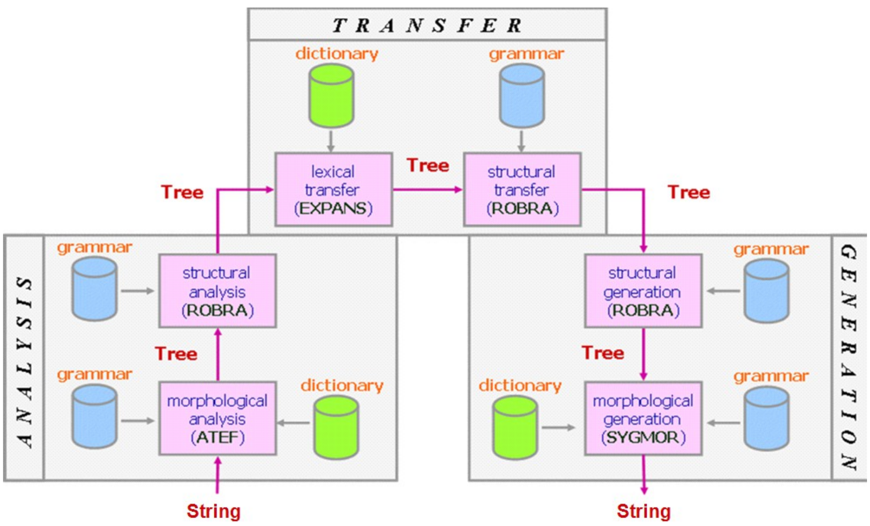

FIGURE 1 – Ariane-G5 phases.

Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pages 50–54, Dublin, Ireland, August 23-29 2014.

2.2 GÉTA's methodology

The approach of the GÉTA group of Grenoble (France), who created Ariane-G5, is a second generation MT, in which the text to be translated is first transformed into an abstract representation, as independent of any language as possible, so this abstract representation can then be translated in any other language. The abstract representation is a multi-level structure (m-structure) ideally containing the logic (predicate-argument) and semantic data that are the most language-independent computed in this approach. As this deep level is not always reached, two other (lower) levels are borne by the m-structure: the syntagmatic level and the syntactic dependency level, so the translation system will output the best it can do.

As one can see in Figure 1, the development is made of modules corresponding to the different steps of the translation. If we concentrate on the analysis (the systems we are working on are X-French and X-English systems so firstly on analysers for the X languages), the work consists in developing monolingual dictionaries containing all the information necessary for the analysis, as well as structural analysers. As such linguistic descriptions are rather complex, one first needs to specify what will be programmed, especially for the structural part. GÉTA's answer to this issue consists in making a list of the different structural phenomena found in the language, each one being represented as a correspondence between a string and its abstract representation ("charts"), and establishing links between the charts so the charts can include references to other charts. One can think it roughly as derivation rules in formal grammars in which terminal elements are classes of words and non-terminal elements are charts. For example, a noun phrase (the string) such as [adjective+noun] can be represented as NP(AP,noun) where AP refers to a chart of general adjective phrases, possibly containing adverbs as in "a very cute cat". The formalism for those charts has initially been called "static grammar" and later SCSG (Static Correspondence Specification Grammar).

3 Parallel work on Khmer, Hindi, Lao and Thai languages

This work aims at elaborating an efficient and simple methodology for developing MT systems for groups of under-resourced languages. We are using for that purpose a small corpus consisting in Saint-Exupéry's *Little Prince* in Khmer, Hindi, Lao and Thai which are our source languages, and our target languages are French and English. Two of the authors, Guillaume de Malézieux and Vincent Berment, are working on Khmer and Lao, as two other persons, Jennifer Wong and Satenik Mkhitaryan, are working on Thai and Hindi.

3.1 Reuse of existing linguistic modules

The systems developed under Ariane-G5 are made of linguistic module dedicated to each step of the translation process (analysis, transfer, generation). In GÉTA's approach, analyses are independent from generations so an analyser for a specific language can be used with a generation of any other language. As French an English modules are available under BSD licence (among many others), we are using them for our work so the analysers and the transfers have to be developed.

3.2 Segmentation and POS tagging

In the case of Khmer, Lao and Thai, one needs to segment into words first, as the writing systems do not include spaces between words. This is done by Motor, a segmenter performing a maximum matching algorithm. It is currently available for Burmese, Khmer, Lao, Thai and Tibetan. Within the limits of our small corpus, the obtained segmentation is 100% correct (the figure reached for general corpora is significantly lower). In order to create the first step called "morphological analysis" in Figure 1, we need a list of words with a number of features that will be used for the analysis. To achieve that, we fill an Excel file with the required data. The following figure is an extract of the Excel file that describe a noun phrase with a possessive attribution. Note that Hindi is not completed and was not included is this paper.

Lao	UL	CAT	Thaï	UL	CAT	Khmer	UL	CAT
ຮູບແຕ້ມ	image	N	รูป	image	N	គំនូរ	image	N
ຂອງ	of	S	ของ	of	S	របស់	of	S
ຂ້ອຍ	I	R	ฉัน	I	R	ខ្ញុំ	I	R

FIGURE 2 – Khmer, Lao and Thai data used in the "morphological analysis"

We used parts of speech often found in GÉTA systems: V verb, N noun, A adjunct, R pronoun, S subordination (preposition, subordinating conjunction and linking word), C coordinating conjunction. In Figure 2, LU stands for Lexical Unit, which is a generalisation of lemma that groups together words deriving from the same base such as build, building, builder, etc. That notion is very useful, for example during transfers where it eases paraphrasing.

The example in Figure 2 is an ideal case where the three languages involved are aligned word for word. When it is not the case, we have different lines for the parts in the different languages that are not aligned and we mark them as "similar" thanks to a colour given to those parts. That is used later when specifying the structural analysers as blocks of words that are not aligned may share common structures (see the next section).

After the Excel file is completed, we can then generate automatically the "morphological analysis" source code written in ATEF language, thanks to a tool we developed for that aim. Note that segmenting and POS tagging have their own dictionaries so a special care is needed to ensure their consistency.

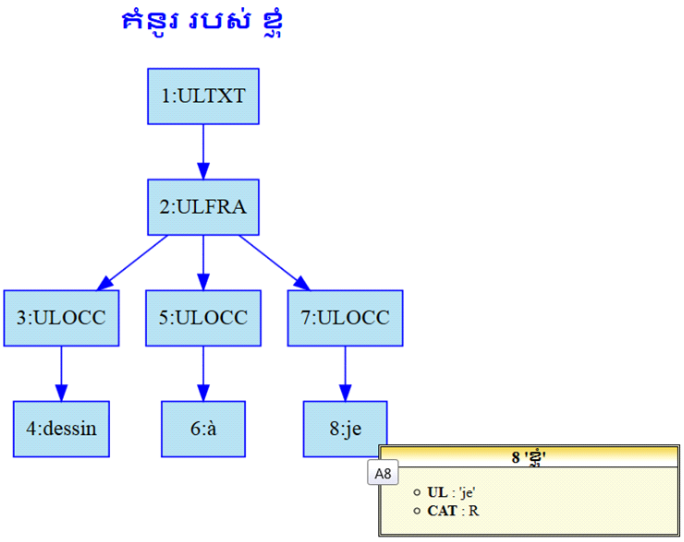

FIGURE 3 – Result of the morphological analysis for គំនូររបស់ខ្ញុំ (Khmer)

3.3 Structural analysis

In order to perform the structural analysis of a text, one needs a formal description of the language. This description, that we call a specification, will be written according to the formalism given by Bernard Vauquois and Sylviane Chappuy [Vauquois et Chappuy, 1985] and mentioned in section 2.2: the static grammars. After we get such specification, we can start programming the analyser in the Ariane-G5 language called ROBRA, which performs tree transformations.

Now let us have a closer look at what a static grammar is like. It is a series of charts, each chart describing a family of strings by associating it to a tree. The charts may refer to each other. For example in order to recognise a complex noun phrase such as "gaz reaction", the two nouns have to be first recognised as separate valid noun phrases (for example, "gaz" is a word that makes sense on its own) so that then they can be gathered into the same tree in order to take a new meaning. So that means the chart describing complex noun phrases refers to the chart describing simple noun phrases. As a consequence, all the charts have to be organised in the grammar so that the ones describing elementary phrases, that are the ones that do not need referring to another chart, come first. Then come the charts describing simple phrases, because they can only refer to lower charts in this hierarchy. At last come the charts for complex groups, they can refer to any chart in the grammar.

Now to write the charts, we need a list of variables to gather all the information we need. They can be of different types, but for the purpose of our study, we will only need basic information. Because we use the limited vocabulary of the *Little Prince*, we won't have to work much on disambiguation. So for now we are only using POS information, with some refinements to recognise mass nouns from countable nouns, and some subcategories of verbs. As an example, we will present the chart describing the possession noun phrases, that are built identically in the three languages: noun + particle "of " + personal pronoun. Here in order to write a chart that could apply to Lao, Thai and Khmer languages at a time, we will use the variable OF to refer to របស់ in Khmer, ของ in Thai, and ຂອງ in Lao. A static chart is divided into three zones. The first one is a string-tree correspondence, describing the structure to be recognised. Each node and leaf receives a number. In FIGURE 4, the root node of our noun phrase is the number 1. Numbers 2, 3 and 4 are the leaves, and each cross below represents a word of the string. The square brackets around number 3 mean that it is optional. The last two lines at the bottom of the tree give information about the words. For example leave 2 is a noun, and more precisely a common noun, leave 3 is a subordinating and its LU is the particle OF, and at last, leave 4 is a personal pronoun. One particularity in this tree is the fact that the node 3 is not linked to the root. This is because although the particle needs to be taken into account during the analysis, we chose not to have it appear into the tree. All the information it carries will be transferred into other nodes. Zone 2 of the static chart provides complementary information on the condition necessary for the structure to be correct. This could be semantic information on one of the nodes, or the presence of one node excluding another, etc. But we do not need any information of this type in the chart we are studying. At last, it is in zone 3 that we present the actions to be taken on the tree. In our case, we store in a variable the possession relation. We also assign the noun of leave 2 to be the governor, that is to say the head, of the phrase.

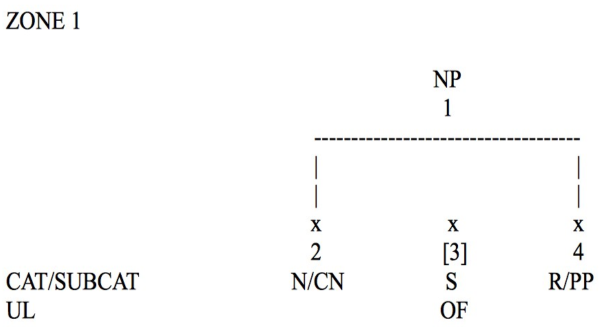

FIGURE 4 – String-tree correspondence

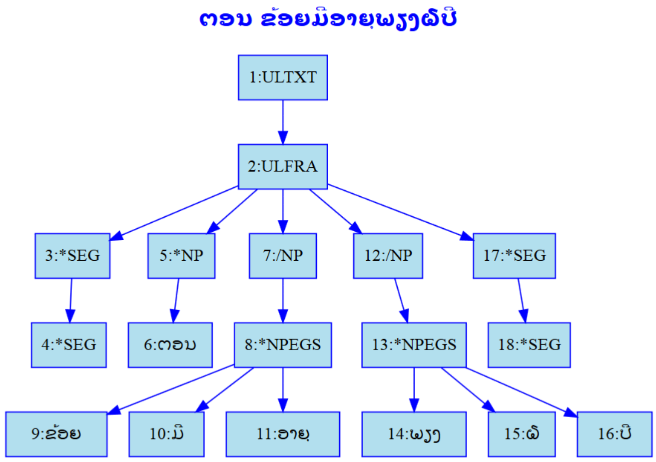

FIGURE 5 – Example of structural analysis for a Lao phrase

3.4 Lexical transfer

In transfers, we transform the Lexical Units and their variables from the source to the target lexical spaces. As we found lexical similarities between Thai, Lao and Khmer languages — ULs are between 50% and 70% common —, a large part of the transfers is also common to those languages.

4 Conclusion

In this paper, we presented an ongoing work. A lot remains to be done but we already observe that working in parallel on several languages brings a lot of advantage. For example, when a question raises on the methodology, on how we can build a specific static chart, etc., people working on any language can answer. For this purpose, the Ariane/Heloise community has set-up a Web site and enriches it continuously: lingwarium.org. Also, as for the structural phases, we noted that many structures were common between Khmer Lao and Thai (Hindi development is late because of the few common features shared with the other languages), thus reducing the effort for making the static grammars. We also noted that the time to develop the transfers were dramatically reduced as a large part of them were common to the three languages. That remains to be further evaluated but we are already convinced it is a way that will help reaching Christian Boitet's prediction that 600 languages will have access to machine translation [Boitet, 2013].

Acknowledgements

We would like to thank Jennifer Wong and Satenik Mkhitaryan for their contribution, as well as Michel Antelme who helped a lot for the work on the Khmer language.

References

Bachut D., Le projet EUROLANG : une nouvelle perspective pour les outils d'aide à la traduction, TALN 1994 Proceedings ,PRC-CHM Days, Marseille University, April 7-8th 1994.

Bachut D., Verastegui N., Software tools for the environment of a computer aided translation system, COLING-1984, Stanford University, pages 330 to 333, July 2-6th 1984.

Berment V., Méthodes pour informatiser des langues et des groupes de langues « peu dotées », PhD Thesis, Grenoble, May 18th 2004.

http://portal.unesco.org/ci/fr/files/16735/10914394223these_Berment.pdf/these_Berment.pdf

Berment V., Boitet C.: Heloise — A reengineering of Ariane-G5 SLLPs for application to π-languages, COLING 2012, Bombay, December 2012

Boitet C., Le point sur Ariane-78 début 1982 (DSE-1), vol. 1, partie 1, le logiciel, ADI Contract report n° 81/423, April 1982.

Boitet C., Guillaume P., Quézel-Ambrunaz M., A case study in software evolution: from Ariane-78.4 to Ariane-85, Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, Hamilton, New York, August 14-16th 1985.

Boitet C., Current machine translation systems developed with GETA's methodology and software tools, Translating and the Computer 8, November 13-14th 1986.

Boitet C., La TAO à Grenoble en 1990, 1980-90 : TAO du réviseur et TAO du traducteur, LATL and CNET, Lannion, 1990.

Boitet C., A research perspective on how to democratize machine translation and translation aids aiming at high quality final output, MT Summit VII, Kent Ridge Digital Labs, Singapour, pages 125 to 133, September13-17th 1999.

Boitet C., A roadmap for MT: four « keys » to handle more languages, for all kinds of tasks, while making it possible to improve quality (on demand), International Conference on Universal Knowledge and Language (ICUKL 2002), Goa, November 25-29th 2002.

Boitet C., Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes, TALN 2008, Avignon, June 9-13th 2008.

Boitet C., Les logiciels traduiront 600 langues dans dix ans, Les dossiers de la Recherche, n°4, June-July 2013.

Chappuy S. Formalisation de la description des niveaux d'interprétation des langues naturelles, Thesis, 1983

Delavennat E., Comparaison des systèmes de décoration des linguiciels traitant les langues FRA, ENG, ALD, RUS, final report, Traouiero project, 2010

Del Vigna C., Berment V., Boitet C., La notion d'occurrence de formes de forêt (orientée et ordonnée) dans le langage ROBRA pour la traduction automatique, Approches algébrique, logique et algorithmique, ATALA, ENST Paris, December 1st 2007.

Collective work, Maquette Pédagogique du BEX FEX, GETA Document, 1983

Guillaume P., Ariane-G5 : Les langages spécialisés TRACOMPL et EXPANS, GÉTA document, June 1989.

Guilbaud J.-P., Ariane-G5 : Environnement de développement et d'exécution de systèmes (linguiciels) de traduction automatique, GDR I3 ATALA, Paris, November 1999.

Tang E.K., Natural languages Analysis in machine translation (MT) based on the STCG, PhD thesis, Sains Malaysia University, Penang, March 1994

Vauquois B., Aspects of mechanical translation in 1979, Conference for Japan IBM Scientific program, July 1979.

Vauquois B., Computer aided translation and the Arabic language, First Arab school on science and technology, Rabat, October 1983.

Vauquois B., Chappuy S., Static grammars, A formalism for the description of linguistic models, Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, Hamilton, New York, August 14-16, 1985.

Zaharin Yusoff, Strategies and heuristics in the analysis of a natural language in machine translation, PhD thesis, Sains Malaysia University, Penang, March 1986.