SMART: Research Directions

Nicola Cancedda (for the SMART Consortium)
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
email: Nicola.Cancedda@xrce.xerox.com

***Abstract*— More than half of the EU citizens are not able to hold a conversation in a language other than their mother tongue, let alone to conduct a negotiation, or interpret a law. In a time of wide availability of communication technologies, language barriers are a serious bottleneck to European integration and to economic and cultural exchanges in general. More effective tools to overcome such barriers, in the form of software for machine translation and other cross-lingual textual information access tasks, are in strong demand. Statistical methods are a promising approach, in that they achieve performances equivalent or superior to those of rule-based systems, at a fraction of the development effort. There are, however, some identified shortcomings in these methods, preventing their broad diffusion. SMART[1] is an attempt to address these shortcomings by deploying the methods of modern Statistical Learning. The scientific focus is on developing new and more effective statistical approaches while ensuring that existing know-how is duly taken into account. Field evaluation on three user scenarios will ensure that advances make their way out of the laboratories, in the form of both improvements over existing technologies and of new applications.**

I. Extended abstract

The multiplication of communication means and their worldwide diffusion make language barriers more critical than ever and create the need for more effective tools, be they in the form of translation aids or of engines for searching multi-language document collections.

Traditionally, Machine Translation (MT) is based on large and complex sets of handwritten rules. Starting from the late 1980s, though, methods for automatically acquiring knowledge from large amounts of manually translated documents have been described and tested [1], [2]. As such knowledge is in the form of parameters of a probabilistic model, these methods are collectively referred to as Statistical Machine Translation (SMT).

In most initial probabilistic models the building blocks were pairs of words of the two languages, but recent work [3], [4], [5], [6] suggests that better results can be obtained if larger building blocks (bi-phrases) are used instead. Currently, the Machine Translation systems performing best in comparative evaluations[2] are phrase-based log-linear probabilistic models.

Despite its success, though, the situation of SMT is still only partially satisfactory:

1) All proposals so far involve training model parameters by maximising some function which either cannot be computed exactly or is characterised by many local maxima. Moreover, all approaches assume that at decoding time a complex heuristic search through a combinatorially large set of alternatives is performed. This in turn requires relying on identifying suitable heuristics with little or no performance guarantees.
2) Current SMT systems are trained in batch mode on a training set, and the corresponding translation model is frozen and used at operation time. This is quite unsatisfactory when Computer-Aided Translation tools are envisaged to improve the productivity of professional human translators, a scenario which seems one of the most realistic for actual deployment of MT technologies. In this context, the human translator is rapidly annoyed by a system that will stubbornly repeat the same incorrect translation for repeated or similar source language material.
3) Probabilistic models strongly depend on the collection of manually translated documents they are trained on. While some large collections of documents with their translations (parallel corpora) are indeed easily available, (Europarl [7], ACQUIS [8], MULTEXT-East, Canadian Hansard, Hong Kong Hansard), it is often the case that no large parallel corpus is available for a given language pair in the domain and genre of interest. In such cases, the system designer is left with the uncomfortable choice of training on a small corpus of appropriate documents or on a larger but less appropriate corpus.
4) Another major obstacle for current SMT systems, and a disadvantage with respect to rule-based systems, consists in the difficulty of producing fluent output. While all SMT systems do incorporate some form of language modelling to ensure that the readability of the target sentence is somehow taken into account, most of them rely on very simple n-gram based models. Such models

[1] Statistical Multilingual Analysis for Retrieval and Translation. The SMART Consortium is composed of:

- Xerox S.A.S.
- Amebis d.o.o, Kamnik
- Celer Soluciones S.L.
- Jozef Stefan Institute
- National research Council - Canada
- University of Bristol
- University of Helsinki
- Università degli Studi di Milano
- University of Southampton
- University College London.

[2] http://www.nist.gov/speech/tests/mt/.

are indeed easy to train and fast to apply at decoding time, but they only consider local phenomena, and sometimes lead to grossly ungrammatical and hardly intelligible output.

5) Cross-Lingual Textual Information Access (CLTIA) is a generic term which covers Cross-Language Information Retrieval (CLIR), Categorization/Clustering (CLC/C) and terminology extraction. In current CLTIA systems, queries and document fragments are generally translated word by word, even though multi-word expressions and other contextual dependencies are often crucial to solve translation ambiguities. This raises the important issue of designing or adapting bilingual lexicon extraction methods, language models and tools used in CLTIA to deal with general, non-contiguous multi-word units.
6) Some recent approaches to CLTIA rely on kernel methods in latent, language-independent concept spaces. Such methods do not aim at precisely identifying word translations, but rather at establishing the most important interlingual concepts in a collection of documents, thus capturing different aspects of the translation process. What is the most appropriate combination of latent, language independent analysis and advanced "surface" translation methods is still an open question.

The focus of SMART will be on addressing these issues. Research directions that will be considered can be stated as follows:

1) Recent advances in Machine Learning show that complex problems such as natural language parsing, that used to be considered highly combinatorial and to require heuristic search, can be formalised as global optimisation problems over convex functions of possibly very many variables [9]. We will follow the same strategy to approach the problem of phrase-based machine translation. A first possibility consists in retaining a model in which relatively complex features are extracted from pairs of a source sentence and a candidate translation, but the relative weight of such features is estimated by maximising margin instead of likelihood or smoothed error. A second, more radical solution, consists in departing completely from the complex features currently used in SMT and adopting a kernel-based approach to implicitly consider as features, and weight independently, all possible co-occurrences of word sequences in the source and in the target language.
2) Not much attention (see [10] for an exception) has been devoted to the development of adaptive SMT models. Within the project we will define and implement a theoretical framework for modelling situations in which the user provides feedback by either accepting the proposed translation or by correcting it.
   This will form the basis for a new class of algorithms for online training of translation models, be they log-linear phrase-based models or large-margin models such as those considered in the previous point. In this new setting, the model is initialised by means of a training set, but undergoes constant evolution as the human translator provides feedback. If this update can be done successfully, then this would lead to a class of computer-assisted translation tools that would be better accepted by actual end users.
3) Domain-specific translation model adaptation is in a similar seminal stage. Hildebrand and co-workers [11] propose IR techniques to select from the training set sentence-pairs whose source is similar to a given test-set, and train only on those. Munteanu and co-workers [12] go further, and propose a classifier for identifying sentences which are translation of one another in a comparable corpus. In the course of the project we will investigate these and other domain-adaptation mechanisms, both for the case in which the original training set is still available and the case in which only the trained model is available, but not the data it was trained on.
4) The main approach followed so far for improving translation fluency consists in taking the syntax of the source sentence (as automatically determined by a parser) into account, in a tree-to-string [13] or even a tree-to-tree [14], [15] mapping. While this can be effective, its success depends on the availability of a syntactic parser, which is a complex resource to create. As an alternative, we propose to investigate methods that do not require explicit syntax modelling. String kernels [16] can for instance leverage long-distance information not available to standard n-gram based language models. Rational kernels [17] can be used to take morphological information into account.
5) We will explore what combinations or variations of latent, language-independent analysis methods are most appropriate for CLTIA tasks. We will also explore to what extent some techniques from phrase-based SMT and multilingual lexicon extraction can be used to support latent methods for CLTIA and, reciprocally, how phrase-based SMT could usefully exploit results from latent, language-independent analysis.
6) The main approach to CLIR used so far consists in using available bilingual resources to translate queries prior to a standard monolingual search [18]. However, most such resources provide word-to-word translation, without any contextual clues, and are not tuned to particular domains. We will develop new methods for extracting multilingual lexica from available, domain-specific corpora, both parallel (multilingual corpora in which the documents are translation of each other) and comparable (multilingual corpora in which documents cover the same topics) that can complement existing resources and be tuned to particular domains. These lexica will contain complex entries to allow translating non-continuous multi-word units.
7) Latent, language independent analysis approaches exploiting the potential of kernel methods have recently been proposed to address most of CLTIA tasks ([19],

[20], [21]). We will explore what is the most appropriate combination of latent, language independent analysis and advanced "surface" translation methods. We will also explore to what extent techniques from phrase-based SMT and multilingual lexicon extraction can be used to support latent methods for CLTIA and vice-versa.

While we believe that a solid theoretical understanding is essential, we will experimentally evaluate the effectiveness of identified solutions. This will be done by deploying our emerging technologies in three different application scenarios: one involving Computer-Aided Translation (CAT) and two involving comprehension aids coupled with Cross-Language Information Retrieval (CLIR).

In the first scenario, the user is a professional translator. Many translators are nowadays comfortable using Translation Memories (TM). When a sentence to be translated is found in a database of previously translated sentences, the corresponding translation is proposed to the user, who can use it as such or edit it. In some cases, a translation can be proposed even if the source sentence is not exactly the same, provided that it is sufficiently similar according to some metrics. We propose to study the integration, in this same framework, of suggestions coming from an MT system. Such scenario is not new, but, besides allowing evaluating the impact of new models and algorithms in terms of improved productivity, it provides an ideal context for studying on-line adaptive algorithms.

In the second scenario, the user is a technical expert (or trouble-shooter), answering telephone calls from customers experiencing some form of problem with a device or software. The user conducts a conversation in language A with a customer, and at the same time accesses technical documentation in a searchable database in language B. The trouble-shooter can be aided by allowing him/her to query the documentation directly in language A. Once, using CLIR techniques, the relevant documents in language B are retrieved, the user can furthermore be provided with comprehension aids, ranging from term translation to full machine translation, that help him/her provide an explanation in language A to the customer.

The third scenario is similar to the second, although in this case the user is a person accessing the popular multilingual WikiPedia on the web. The user can enter a query in language A. Relevant WikiPedia entries in several languages are returned, and comprehension aids are then provided to help the user access the content of those entries in languages he/she is not familiar with.

The second and third scenarios above provide a testbed for several aspects of the research planned in the project, including domain adaptation of translation and language models, Cross-Language Information Retrieval, bi- and multi-lingual lexicon extraction.

REFERENCES

[1] F. P. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, 1990.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, 1993.

[3] F. J. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *Proc. of the joint conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, 1999.

[4] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[5] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of the Conference on Human Language Technology*, Edmonton, Canada, 2003.

[6] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[7] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," draft, unpublished, available as http://people.csail.mit.edu/ koehn/publications/europarl.ps.

[8] T. Erjavec, C. Ignat, B. Pouliquen, and R. Steinberger, "Massive multilingual corpus compilation: Acquis and totale," in *Proc. of the Second Language and Technology Conference*, Poznan, Poland, 2005.

[9] B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning, "Max-margin parsing," in *Proc. of the COnference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[10] L. Nepveu, G. Lapalme, P. Langlais, and G. Foster, "Adaptive language and translation models for interactive machine translation," in *Proc. of the Conference on Empitical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[11] A. S. Hildebrand, M. Eck, V. Stephan, and A. Waibel, "Adaptation of the translation model for statistical machine translation based on information retrieval," in *Proc. of the Meeting of the European Association for Machine Translation (EAMT)*, Budapest, hungary, 2005.

[12] S. Munteanu, A. Fraser, and D. Marcu, "Improved machine translation performance via parallel sentence extraction from comparable corpora," in *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004)*, Boston, Massachussets, 2004.

[13] K. Yamada and K. Knight, "A syntax-based statistical translation model," in *Proceedings of the annual meeting of the Association for Computational Linguistics (ACL)*, Toulouse, France, 2001.

[14] D. Wu, "Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammars," in *Parallel Text Processing: Alignment and Use of Translation Corpora*, J. Veronis, Ed. Dordrecht, The Netherlands: Kluwer, 2000.

[15] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, 2005.

[16] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *Journal of Machine Learning Research*, vol. 2, Feb 2002.

[17] C. Cortes, M. Mohri, and J. Weston, "A general regression technique for learning transductions," in *Proc. of the International Conference on Machine Learning*, 2005.

[18] J. Savoy, "Report on clef-2003 multilingual tracks," in *Comparative Evaluation of Multilingual Information Access Systems, CLEF*, ser. Lecture Notes in Computer Sciences, C. Peters, J. Gonzalo, M. Braschler, and M. Kuck, Eds. Springer Verlag, 2004, vol. 3237.

[19] Y. Li and J. Shawe-Taylor, "Using KCCA for japanese-english cross-language information retrieval and classification," *Journal of Intelligent Information Systems*, 2005, to appear.

[20] B. Fortuna and J. Shawe-Taylor, "The use of machine translation tools for cross-lingual text mining," in *Proc. of the ICML Workshop on Learning with Multiple Views*, 2005.

[21] W. Buntine and A. Jakulin, "Discrete component analysis," in *Subspace, Latent Structure and Feature Selection*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 3940.