Hybrid machine translation.

Applications of Machine Translation

Hybrid machine translation takes advantage of the strengths of statistical and rule-based approaches. Typically, the methodology of one of these approaches is used in order to improve the quality of translation of other approaches. For example, statistical methods can be used to select one of several possible translation meanings for a word, while an entire sentence can be translated based on grammatical rules. Alternatively, grammar rules are used to present the results of statistical computer translation in grammatically correct form.

MT sources of information can be rules or data. The former is linguistically motivated, and the latter is more statistically motivated. Rules. MT approaches based on rules (i.e. RBMT) use linguistic information such as monolingual and bilingual dictionaries combined with human linguistic knowledge. Rules are developed manually to transfer text in a source language text into a target language text. Most popular RBMT approaches apply three different phases: analysis, transfer and generation. Data. Data-driven MT approaches use information from data and complex algorithms which together are capable of modeling translation. Data driven MT includes: example (EBMT) and statistical-based (SMT). By definition, EBMT approaches perform a direct translation by analogy and it can be seen as a pattern matching problem. Unlike these, SMT systems try to find the most probable translation given the source sentence, by reference to the models built using data such as the translation and language model (Brown et al., 1993). SMT can be classified into phrase, syntax and hierarchical. The main difference among these models is the structure of the bilingual units which can be built from: (1) plain text in the case of phrase models; (2) more complex data including grammars and dependency trees in syntax models; and (3) plain text but allowing hierarchical units in hierarchical systems. Given that hybridization is the focus of this study, we will consider this latter criterion (sources of information) in order to distinguish MT paradigms. Within this category, we detail a wide variety of hybridization approaches.

Hybridization of machine translation architectures

Several different methodologies have been used to hybridize MT within and across paradigms. As shown in Fig. 1, hybridization of RBMT and corpus-based MT can be classified into those guided by RBMT or guided by corpus-based MT. The former integrates data information into a rule-based architecture; the latter integrates linguistic rules into a corpus-based architecture.

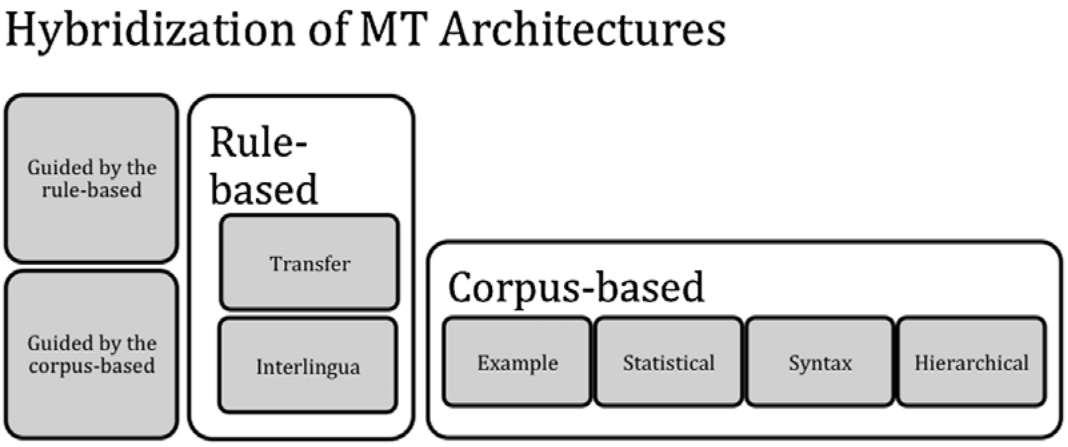

Figure 1. Classification of hybrid MT architectures

Hybridization guided by RBMT

There are several kinds of strategy within this category: introducing a corpus to build the RBMT system, introducing corpus-based tools to weight the RBMT output and carrying out a statistical post-editing of a RBMT output.

Using a corpus to build the RBMT system. The main reason for using data when building a RBMT system is to reduce its cost and the time and effort required. A quite straightforward approach is to enhance dictionaries with phrases or examples extracted from parallel corpora and extract new entries from Babel Net and Wiktionary. More complex approaches extract transfer rules, build lexical selection modules using parallel corpora with finite-state transducers or Maximum Entropy Markov Models, and combine several of these techniques.

Corpus-based tools for weighting the RBMT output. There is work that focuses on improving the RBMT output by integrating tools such as language models or stochastic parsers. Papers like [2] show a hybrid translation system guided by the RBMT engine and, before transference, a set of partial candidate translations provided by SMT subsystems is used to enrich the tree-based representation. The final hybrid translation is created by choosing the most probable combination among the available fragments with a statistical decoder in a monotonic way (see Fig. 2). In addition, there are RBMT systems that introduce machine learning techniques such as classifiers in order to identify the set of appropriate translation candidates. Recent experiments by Systran build a statistical inference module to replace the RBMT transfer module and experiments by Lingenio show that RBMT systems can learn morphological classification, semantic and syntactic information from corpus data.

Statistical post-editing of RBMT outputs. There are studies that carry out statistical post-editing for RBMT systems and it is even a commercial reality as pointed out in [3]. Generally speaking, these approaches consider RBMT outputs as source sentences and post-edited results as target sentences. In other cases, [4] confidence estimation measures are used instead of manually post-edited results. The statistical module tends to be implemented with Moses. In this case, RBMT and SMT paradigms are concatenated but not integrated at the architecture level.

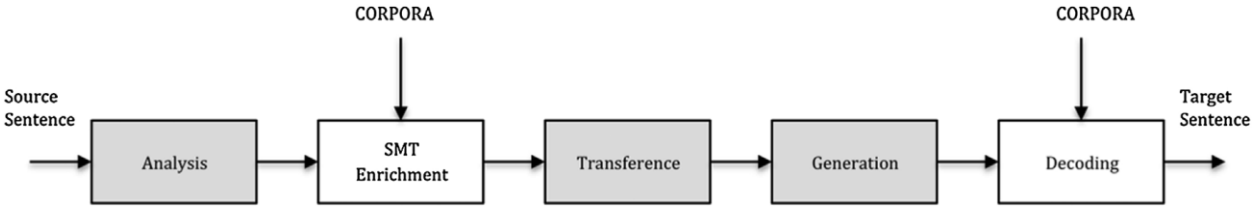

Figure 2. Schema of hybridization guided by RBMT

Hybridization guided by corpus-based MT

A hybrid system guided by corpus-based MT may incorporate rules or just combine various corpus-based MT approaches. There are basically two main ways of integrating rules into corpus-based MT approaches: using rules at pre/postprocessing and integrating dictionaries/rules into the core model.

Rules at pre/post-processing. Pre-processing rules have been used to reorder the source sentence into a form that better matches the target language. The schema for this type of strategy is shown in Fig. 3. Post-processing rules for morphology generation have been introduced by means of a combination of machine learning and the introduction of dictionaries.

Finally, a set of both pre-processing and post-processing rules have been compiled ad-hoc for the Spanish-Catalan translation pair in [5], in order to solve the normalization problems typically found in noisy corpora.

Incorporating dictionaries/rules into the core model. Rules may be integrated into the core model of corpus-based MT approaches. Early work such as [6] integrates morphology and syntax knowledge from the RBMT system dynamically into an EBMT system. In other cases, RBMT systems have been integrated into the phrase-based SMT modules. For example, in [7] is used RBMT information to improve statistical word alignment. Then, Eisele et al. [8] augment the standard phrase table with entries obtained after translating the data with several RBMT systems. The resulting phrase table thus combines statistically gathered phrase pairs with phrase pairs generated by linguistic rules. Similarly, Sánchez-Cartagena et al. [9] enrich the phrase table with bilingual phrase pairs matching transfer rules and dictionary entries from a shallow-transfer RBMT

system and carrying out a comparison with an earlier paper. Further work by these latter authors integrates a commercial RBMT system with a hierarchical SMT system by extracting rules from RBMT translations. The hybrid system inherits the lexicons from both sub-systems as well as local syntactic constructions defined in RBMT. From a different perspective, Ahsan et al. [10] focus on integrating local and long reorderings as well as the generation module from an RBMT system, into the core translation model of a standard statistical system. Furthermore Enache et al. [11] introduce rules from a grammar formalism into the phrase table, and Okuma et al. [12] introduce dictionaries into the phrase table to reduce the number of unknown words.

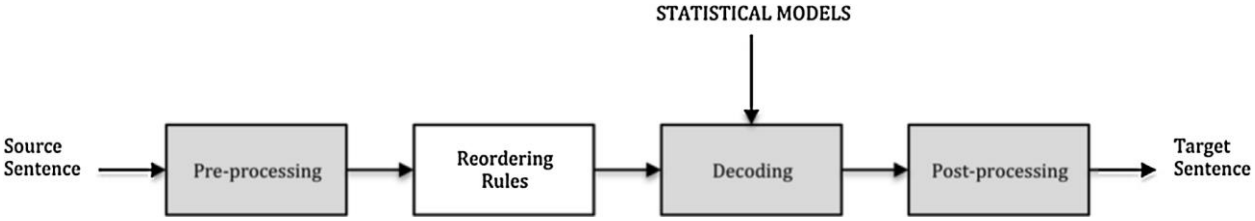

Figure 3. Schema of hybridization guided by corpus-based MT

Hybridization within corpus-based approaches. When combining corpus-based approaches, Groves and Way [13] mix sub-sentential alignments from phrase-based SMT and EBMT systems, proposing to build a hybrid 'example-based' SMT system incorporating marker chunks and SMT sub-sentential alignments. There is an extensive body of work on incorporating translation memories (TM) into phrase-based SMT systems. TM are simply large databases of translated words and sequences of words, generally created by human translators. One of the most recent studies proposes integrated models to make maximum use of TM information during decoding. The aim is to keep all its possible corresponding target phrases for each TM source phrase. The integrated models then consider all corresponding TM target phrases and SMT preferences during decoding. Therefore, the proposed integrated models combine SMT and TM at a deep level. A traditional way that cannot be neglected is the use of templates, which themselves can be considered to be stochastically-extracted transduction type rules. There are also approaches that combine n-gram and phrase SMT in series. The former pre-reorders the source sentences and offers a reordering graph that the latter translates using monotonic decoding.

Finally, there are approaches that are exempt from the requirement for parallel corpora or resources in general. There is an MT method that needs no parallel text and relies on a translation model built from a bilingual dictionary, and a decoder for long-range context. In the same direction, other systems use low resources and a methodology designed to facilitate rapid creation of the MT system for unconstrained language pairs.

Machine translation applications with hybrid components

Among the variety of MT applications, we can name popular ones such as speech translation, cross-lingual information retrieval and computer-aided translation. Hybridization within these applications has been used in different ways and we offer comments some of them without aiming to be exhaustive. See Fig. 2 for a short summary of references.

Speech translation. Frequently, speech translation is addressed as a concatenation of a speech recognizer, a machine translator and a speech synthesizer. Hybridization in this application can be placed in any of the three systems. In speech recognition, hybridization has been done by incorporating neural network approaches into state-of-the-art continuous speech recognition systems based on hidden Markov models (HMMs). There is also the combination of hidden Markov models (HMMs) and learning vector quantization (LVQ), or the use of Support Vector Machines (SVMs) for classification by integrating this method into a HMM-based speech recognition system. In text synthesis, the hybridization has been done by combining concatenative synthesis and statistical synthesis.

Cross-lingual information retrieval. Normally, the application of cross-lingual information retrieval is done by concatenating MT and information retrieval. For example, Mittal et al. [14]

present a hybrid information system combining: (1) an ontology for the retrieval of user's context (2) a user profile that is temporarily updated according to user's browsing behavior and (3) collaborative filtering for considering recommendations of similar users. Elsewhere, Rose and Belew [15] use a combination of symbolic and connectionist artificial intelligence techniques.

Computer-aided translation. Finally, computer-aided translation is by definition a combination of the roles of both man and machine. Recent work uses a machine-aided translation system, which is a hybrid system that applies not only TM technology but also MT methodologies, including the annotation schema of Translation Corresponding Tree (TCT) in the representation of bilingual examples, and the language formalism of Constraint-based Synchronous Grammar (CSG) in analyzing the syntactic structure between the languages; Yamabana et al. [16] also propose a hybrid interactive MT method that combines rule and example-based MT approaches with an interactive man-machine interface. Advanced work in the field such as [17], includes approaches to incremental training or active learning which are representative of live human-machine hybridization where the MT system learns and improves based on human interaction.

References

1. Costa-Jussa, M. R. and J. A. Fonollosa. Latest trends in hybrid machine translation andits applications. Computer Speech & Language, 32(1):3–10, 2015.

2. Labaka, G., Espa na-Bonet, C., Màrquez, L., Sarasola, K., 2014. A hybrid machine translation architecture guided by Syntax. Mach. Transl. 28,1–35.

3. Béchara, H., Rubino, R., He, Y., Ma, Y., Genabith, J., 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In: Proceedings of International Conference on Computational Linguistics (COLING), pp. 215–230.

4. Suzuki, H., 2011. Automatic post-editing based on SMT and its selective application by sentence-level automatic quality evaluation. In: Proceedings of the 13th Machine Translation Summit (MT Summit XIII), International Association for Machine Translation, pp. 156–163.

5. Farrús, M., Costa-jussà, M.R., Mari no, J., Poch, M., Hernández, A., Henríquez, C., Fonollosa, J., 2011. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair. Lang. Resour. Eval. 45 (2), 181–208.

6. Carl, M., Pease, C., Iomdin, L., Streiter, O., 2000. Towards a dynamic linkage of example-based and rule-based machine translation. Mach. Transl.15 (3), 223–257.

7. Hua, W., Haifeng, W.,2004. Improving statistical word alignment with a rule-based machine translation system. In: Proceedings of the 20thInternational Conference on Computational Linguistics. Association for Computational Linguistics, p. 29.

8. Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., Chen, Y., 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In: Proceedings of the 3rd Workshop on Statistical Machine Translation (WMT), pp. 179–182.

9. Sánchez-Cartagena, V.M., Sánchez Martínez, F., Pérez Ortiz, J.A., et al., 2011. Integrating shallow-transfer rules into phrase-based statistical machine translation. In: Machine Translation Summit.

10. Ahsan, A., Kolachina, P., Kolachina, S., Sharma, D.M., Sangal, R., 2010. Coupling statistical machine translation with rule-based transfer and generation. In: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas.

11. Enache, R., Espana-Bonet, C., Ranta, A., Màrquez, L., 2012. A hybrid system for patent translation. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT), May, Trento, Italy, pp. 269–276.

12. Okuma, H., Yamamoto, H., Sumita, E., 2008. Introducing a translation dictionary into phrase-based SMT. IEICE Trans. Inf. Syst. E91-D (July (7)), 2051–2057.

13. Groves, D., Way, A., 2005 Dec. Hybrid data-driven models of machine translation. Mach. Transl. 19 (3–4), 301–323.

14. Mittal, N., Nayak, R., Govil, M.C., Jain, K.C., 2010. Evaluation of a hybrid approach of personalized web information retrieval using the FIRE dataset. In: Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India, New York, NY, USA, pp. 52:1–52:6.

15. Rose, D.E., Belew, R.K., 1989. Legal information retrieval a hybrid approach. In: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ICAIL'89, ACM, New York, NY, USA, pp. 138–146.

16. Yamabana, K., Kamei, S., Muraki, K., Doi, S., Tamura, S., Satoh, K., 1997. A hybrid approach to interactive machine translation: integrating rule-based, corpus-based, and example-based method. In: Proceedings of the 15th International Joint Conference on Artifical Intelligence, IJCAI'97, vol. 2, San Francisco, CA, USA, pp. 977–982.

17. Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., Germann, U., 2014. The MATEC, tool. In: Proceedings of COLING 2014, the 25thInternational Conference on Computational Linguistics: System Demonstrations, August, Dublin, Ireland, pp. 129–132.