Squib

How much hybridization does machine translation need?

Marta R. Costa-jussà[*]
Institute for Infocomm Research,
Singapore

Rule-based and corpus-based Machine Translation (MT) have coexisted for more than 20 years. Recently, boundaries between both paradigms have narrowed and hybrid approaches are gaining interest from the academic and business point of view. However, since the hybrid approaches involve the multidisciplinary interaction of linguists, computer scientists, engineers and informatics, there arise a variety of questions.

While statistical methods currently dominate research work in statistical MT, most commercial MT systems are technically hybrid systems. The research community should more actively investigate the benefits and questions surrounding hybridization of MT systems. This squib discusses about different issues related to hybrid MT including the origins, the architectures, the achievements and frustrations. Understanding hybridization in the wide sense, both rule-based and corpus-based MT systems have benefited from hybridization when correctly integrated. In addition and to some extension, most of the current rule/corpus-based MT approaches are already hybrid since they tend to include some statistics/rules at some stage.

1. Introduction

Hybrid Machine Translation (MT) has gained popularity since the development of first corpus-based MT systems. Some years from now, most commercial MT developers claim to have hybrid MT systems (see conferences as AMTA 2004 and MT Summit 2005); big projects such as EuroMatrix or T4ME develop some research on hybrid MT systems; and popular shared tasks include system combination of MT systems of different nature. Moreover, most production statistical MT systems use rule-based components for things like transliteration and the translation of numbers. So deployed statistical MT systems are in some sense all hybrid systems, with purely statistical systems really only being seen in research work.

The very first publications in hybrid MT (involving corpus-based systems) date from the early 90s with the integration of statistical information in rule-based MT systems (Doi and Muraki 1992). Other pioneer hybridizations include the combination of translation memories with rule-based MT (Heyn 1996); combination of example-based MT with rule-based MT and combination of translation memory with example-based

---

[*] Currently, visiting Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City
E-mail: martaruizcostajussa@gmail.com

MT (Carl and Hansen 1999). More advanced hybrid techniques include the combination of rule-based and several types of corpus-based approaches (Carl et al. 2008). Finally, most recent works on hybridization tend to involve statistical systems with linguistic knowledge (Costa-jussà and Farrús 2014), which reflects that modern statistical MT may be more rule-based than most people realize. Although most probably it is still not readily accepted by the statistical MT community, when using significant linguistic resources, statistical MT systems could be considered hybrid systems (only if broading the concept of hybridization, as it will be discussed later).

Hybridization tends to be guided by one single approach, which is capable of borrowing strengths of other approaches that are actually weaknesses for the guiding one (e.g. a rule-based number translation module can help standard statistical MT to translate numeric tokens like 1,234,567.89, and solving the weakness that statistical MT would be only able to translate numbers seen in the training corpus).

In this squib, we review the conceptual foundation underlying hybridization, contrasting MT approaches from different nature. We report references to several works that help to define and exemplify this MT hybridization. Also we examine the successes and failures of the performance of such systems and propose several future research lines. At the end of this discussion, we should be able to have a clearer idea of what is exactly a hybrid MT system and if MT does really need hybridization.

The remainder of this discussion is organized as follows. Next Section reports the main concepts behind rule-based and statistical MT. Following Section 3 describes several successful hybrid architectures. Section 4 discusses a variety of relevant properties of hybridization that will help to characterize it more in detail. In particular, the properties are degree, effectiveness, complexity, modeling and impact. Final Section 5 reports some interesting hybridization directions that may be explored in the future.

2. Rule-based and Statistical MT main concepts

MT has historically been faced from different perspectives: either following a linguistic or a data-based approach. As follows, we briefly describe one of the most representative systems for each branch that are popularly co-existing nowadays (i.e rule-based and phrase-based statistical systems).

On the one hand, rule-based MT provides a translation based on linguistic knowledge. The translation process is divided in: analysis, transfer and generation. Analysis and generation cover mainly the morphological variations of the languages, the transfer phase is in charge of the grammatical aspects (Hutchins and Sommers 1992).

On the other hand, statistical MT uses probabilistic models to learn translations (Brown et al. 1993). Given a source string s, the goal is to choose the string with the highest probability among all possible target strings t. Phrase-based translation models (Koehn, Och, and Marcu 2003), are estimated directly from aligned bilingual corpora by considering relative frequencies. Modern systems build translations using a decoding model consisting of the log-linear combination of multiple feature functions (Och 2003). Decoding models typically use a small number of features, but there has been some work investigating using a large number of sparse decoding model features, e.g. (Liang et al. 2006; Chiang, Knight, and Wang 2009; Green et al. 2013). A critical model in this combination is the translation model, as it defines the scored transfer rules (or translation units) for converting words and phrases (possibly non-linguistic sequences of words) from one language into another. In addition to the translation model, statistical MT systems use a target language model, which is usually formulated as a probability distribution over strings that attempts to reflect how likely a string occurs

inside a language. Additional feature functions are also introduced in the statistical MT framework to improve the translation, like for instance lexical re-orderings and graphs (Su et al. 2010). Among other advancements that allow statistical MT systems to model more complex linguistic phenomenon, there is the popular hierarchical MT approach (Chiang 2007).

Because rule-based systems explicitly attempt to process every part of the input, they are not subject to some of the pathological behavior of statistical systems such as the occasional random deletion of important content words. However, they exhibit weaknesses in the lexical selection transfer. Differently, statistical MT systems that require plenty of parallel data are more robust and always produce output, which tends to be more natural given the role of the language model. But such data-driven MT systems base their knowledge on bilingual aligned corpora, and the accuracy of their output depends heavily on the quality and the size of these corpora. Large and reliable bilingual corpora are unavailable for many language pairs. In addition, translating into morphologically rich target languages makes the training of data-driven systems even more difficult (due to data sparseness).

3. Hybrid Rule-based and Statistical MT Architectures

The boundaries between rule-based and statistical MT approaches have narrowed and, recently, some approaches have been proposed for constructing hybrid MT systems with a great potential of improvement.

In general terms, hybrid architectures can be defined in the wide sense or in the restricted-sense. The latter would define a hybrid system, which puts together the rule and statistical core architectures. For example, one natural way would be to induce statistical transfer-rules with the corresponding probabilities and integrate them in a rule-based MT system. We are not aware of such a system. However, if we move towards a more wide sense definition of hybrid MT system, there is a broad spectrum of proposed approaches: statistical MT models augmented with linguistic information (including morphological, syntactic or semantic information); or rule-based MT systems using corpora to improve results by enriching their lexicons and grammars and applying new methods for disambiguation.

3.1 Hybrid systems guided by the statistical approach

There is an extensive research regarding statistical MT models augmented with morphological, syntactic and semantic information. Basically, the approaches reported in this Section are mostly phrase-based with some kind of linguistic information, being examples of hybrid MT in the wide sense (i.e. they do not specifically integrate any type of rule-based module into the statistical engine).

Statistical MT models enhanced with morphological information typically include: (1) segmentation approaches, which tend to make vocabularies from the source and target language more similar by separating morphemes from the root part of the word (Habash and Sadat 2006); (2) generation approaches, which assume that not all morphological word forms are present in the training corpus and propose post-processing techniques to generate new word forms (Clifton and Sarkar 2011); and (3) enriching approaches, which tend to include additional information to the low inflected language to make it more similar to the high inflected language (Avramidis and Koehn 2008).

Statistical MT models can also be enhanced with syntactic information. In the widest sense, we may argue whether statistical MT systems that use syntactic knowl-

edge as parse trees (Quirk, Menezes, and Cherry 2005; Zollmann and Venugopal 2006) could be moved towards the category of hybrid systems. In fact, given that models are learned statistically, most of the MT community probably would not agree with this claim. However, many of the reordering systems introduced into statistical MT have been actually analogous to older rule-based systems, at least for the ones developed by sitting down and writing transfer rules over parse trees (e.g. the phrase-based statistical MT systems extended with some pre-reordering manually-extracted rules (Collins, Koehn, and Kučerová 2005)).

Finally, there are different studies that enhance statistical MT with semantic information. There are systems, which include lexical semantics (España-Bonet, Giménez, and Màrquez 2010), shallow semantic knowledge (Wu and Fung 2009) and systems that try to go towards an abstract meaning representation[1].

3.2 Hybrid systems guided by the rule-based approach

Rule-based MT systems have successfully integrated statistical models in different ways. Dove, Loskutova, and de la Fuente (2012) use rule-based MT to produce a first output and then, refines this output through comparison against a language model. Tyers, Sánchez-Martínez, and Forcada (2012) use statistical lexical semantics to disambiguate translation inputs in the source. Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez (2015) generate transfer rules (without probabilities) from a parallel corpus. These rules can be then used in combination with manually written transfer rules if available (Costa-jussà and Centelles 2015).

4. Dimensions of hybridization

In this Section, we are defining and discussing: the degree, effectiveness, complexity, modeling and impact, as general dimensions of hybridization. The objective of such analysis is to better enclose the concept of hybrid MT.

Degree. Hybridization is quite an open concept. We may discuss to use it in a restricted or wide manner or simply decide to speak of different degrees of hybridization.

One extreme position would be to argue that pure rule-based or statistical MT systems do not exist. On the one hand, most of the rules, which constitute the core of rule-based MT systems, are simply some formulation of statistical knowledge (i.e. human generalization of observations). On the other hand, statistical MT systems always use rules at some point (i.e. preprocessing as tokenization, translation of numbers or transliteration).

The other extreme position of viewing things is that all systems are pure because they are using rules. Some systems are using simplistic rules while others are using more sophisticated rules. Also, some systems are based on rules that are automatically learned from corpora using statistical machine learning techniques, while other systems are based on hand-coded rules. And in the past, machine-learning based systems have tended to learn rather simplistic rules as compared to handcrafted systems.

In between the two previous positions, there is the option of hybridization degrees. On the one side, there would be the systems with very low degrees of hybridization, which would correspond to the standard rule-based and statistical MT systems. On the

---

[1] Refer to the main web page describing the English sentences paired with readable semantic representations at http://amr.isi.edu/
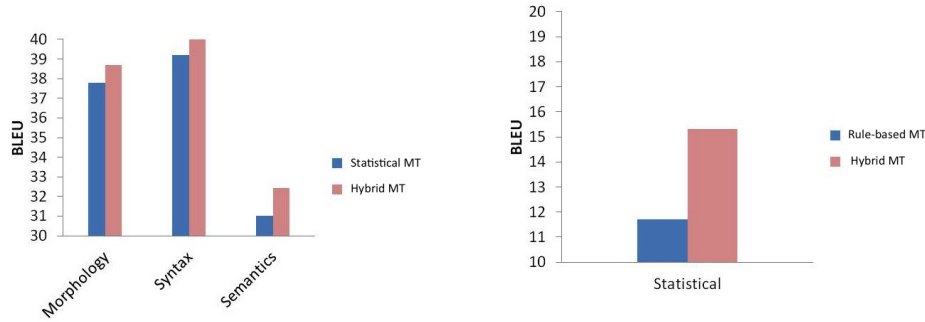
Figure 1: (Left) Impact of using morphological, syntax and semantic linguistic knowledge in a statistical-based MT system in terms of BLEU. (Right) Impact of using statistical knowledge in a rule-based MT system in terms of BLEU.

other side, there would be the systems with high degrees of hybridization, which would include a mixture of architectures either at the level of transfer-rules or decoding. In the middle, there would be a high range of approaches using statistical and linguistic knowledge or rules and statistics.

Effectiveness. More important than deciding if systems are defined to have a low or high level of hybridization, it is to determine if hybridization improves translation. As mentioned in the previous Section, there are a lot of approaches, which prove that the translation quality improves when using linguistic knowledge in statistical MT, and the other way round, when using statistics in rule-based MT.

Figure 1 (left) shows how a phrase-based statistical MT can be improved when enhancing different linguistic levels in the translation task from Arabic-to-English. The translation measure is percentage BLEU (Papineni et al. 2002). The morphology approach is the one by Habash and Sadat (2006), which is trained the system on 5 million words and tested using the NIST 2005 evaluation test set. The approach consists of segmenting the Arabic to make it morphologically more similar to English. The syntax approach is the one by Carpuat, Marton, and Habash (2010), which is trained the system on 64 million words and tested on the NIST 2008 evaluation test set. The approach consists of making a pre-reordering of post-verbal subjects. Finally, the semantic approach is the one by España-Bonet, Giménez, and Màrquez (2010), which is trained the system on 5 million words and tested on the NIST 2008 evaluation test set. The approach consists in incorporating a local discriminative phrase selection model to address the semantic ambiguity of Arabic.

Figure 1 (right) shows some results of using statistical knowledge in rule-based MT systems. A successful approach is by España-Bonet et al. (2011) were the rule-based system is enriched with a set of partial candidate translations provided by statistical MT subsystems. The final translation uses a decoder to choose the most probable combination. BLEU in Spanish-to-Basque News task improves around 4 points.

Complexity vs. improvement. Up to here, we have shown that hybridization can benefit both rule-based and statistical MT. Could we generalize the popular sentence of "there is no data like more data" into "there is no hybridization like more hybridization"? Unfortunately, none of the two sentences do seem to be correct. And definitely, it

would not be true that the higher the level of hybridization, the better the system. When it comes to increasing the amount of training data, there are more parameters that play their role: the quality and domain of data (Koehn and Haddow 2012), among others. When it comes to increasing the amount of hybridization, we could cite some cases where adding linguistic knowledge penalizes the statistical MT system as proven when limiting translation units (Koehn, Och, and Marcu 2003) or reordering (Costa-jussà et al. 2007) with linguistic criteria. Also in the case of rule-based MT, using statistical data can be harmful, i.e. in general manual transfer rules still achieve better results compared to the automatically extracted ones (Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez 2015).

Impact. It is fairly difficult to measure the impact of hybrid MT in terms of works published. We have used the *mt-archive* resource[2] and we have counted all papers in statistical, rule-based and hybrid MT just to have a vague idea of the publication directions. We have seen that since 2010 the proportion is 80%, 5% and 15%, respectively. These results make hybrid MT slightly more appealing than rule-based MT but still very far from the interest that statistical MT rises.

5. Future research lines

There is a lot of work to be done towards a higher level of hybridization. On the one hand, inspired in previous promising works (España-Bonet et al. 2011), we have identified some ways of building a restricted hybrid architecture given a rule-based MT system and available parallel and monolingual corpora (see Figure 2). First, starting with the core of a rule-based system, there is the necessity of extracting transfer-rules from parallel corpus and computing a translation probability to each rule. This would allow to building rule-based MT systems by a monolingual human linguist. At the moment, rule-based MT systems have to be developed by bilingual linguists or at least people who are proficient in the source and target language. Second, in order to help rule-based MT systems be more fluent and natural, it would be nice to integrate a language model in the generation step. The language model could be n-gram-based, syntax-based or trained using deep learning. In each case, a different decoding would be required to be integrated in the system. And third, additional feature functions as the popular lexical ones or others that introduce source context information can be used together with the above language model.

On the other hand, inspired again by promising techniques (Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez 2015), we have identified some ways of building a restrict hybrid architecture given a statistical-based MT system and some additional linguistic resources. First, starting with the core of a statistical-based system, there is the necessity to integrate transfer-rules and find the best way to have them coupled with the translation model. Deep transfer-rules can be added to deep-statistical systems like syntax or hierarchical, making sure to adopt some compatible grammar formalism. In this research line, it is important to identify the best way to make the integration: giving priority to the transfer-rules over the statistical ones or extracting a score to make transfer-rules compete in equal conditions to the statistical ones. And second, widening hybridization, other approaches to improve statistical MT systems at the level of morphology may go in the direction of works like Clifton and Sarkar (2011) moving

---

2 Refer to the web page: http://www.mt-archive.info, which aims to cover comprehensively English-language publications since 1980
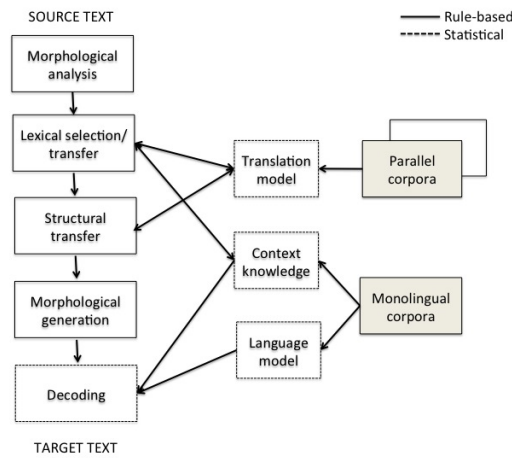
Figure 2: Block diagram of the hybrid MT architecture.

towards a non-pipelined approach. At the level of syntax, new grammar formalisms may be experimented as well as alternative extensions of hierarchical systems. Finally, in the field of semantics, vector-space projections, space reductions and deep learning open new horizons.

These are encouraging research directions that may lead us to a more natural coupling of linguistics and statistics. There are a lot of remaining questions to be solved including the right way of implementing the mentioned research. However, the perspectives are promising and without doubt great advances in MT may arise from multidisciplinary collaborations. Ultimately, this is what hybridization is all about.

References

Avramidis, Eleftherios and Philipp Koehn. 2008. Enriching morphologically poor languages for Statistical Machine Translation. In *Proc of the ACL-HLT*, pages 763–770, Columbus.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of Statistical Machine Translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Carl, Michael and Silvia Hansen. 1999. Linking translation memories with example-based machine translation. In *Proc. of the MT-Summit*, pages 617–624, Singapore.

Carl, Michael, Maite Melero, Toni Badia, Vincent Vandeghinste, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, and Olga Yannoutsou. 2008. METIS-II: low resource Machine Translation. *Machine Translation*, 22(1-2):67–99.

Carpuat, Marine, Yuval Marton, and Nizar Habash. 2010. Improving Arabic-to-English Statistical Machine Translation by reordering post-verbal subjects for alignment. In *Proc of the ACL*, pages 178–183, Uppsala.

Chiang, David. 2007. Hierarchical Phrase-based Translation. *Comput. Linguist.*, 33(2):201–228.

Chiang, David, Kevin Knight, and Wei Wang. 2009. 11,001 new features for Statistical Machine Translation. In *Proc. of HLT-NAACL*, pages 218–226, Boulder.

Clifton, Ann and Anoop Sarkar. 2011. Combining Morpheme-based Machine Translation with post-processing morpheme prediction. In *Proceedings of the ACL HLT*, pages 32–42, Portland.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for Statistical Machine Translation. In *Proc. of the ACL*, pages 531–540, Ann Arbor.

Costa-jussà, Marta R. and Jordi Centelles. 2015. Description of the Chinese-to-Spanish Rule-based Machine Translation system developed with a Hybrid combination of Human Annotation and Statistical Techniques. *ACM Transactions on Asian Language Information Processing*, Submitted.

Costa-jussà, Marta R., Josep M. Crego, Patrik Lambert, Maxim Khalilov, José A. R. Fonollosa, José B. Mariño, and Rafael E. Banchs. 2007. Ngram-based Statistical Machine Translation enhanced with multiple weighted reordering hypotheses. In *Proc of the WMT*, pages 167–170, Prague.

Costa-jussà, Marta R. and Mireia Farrús. 2014. Statistical Machine Translation enhancements through linguistic levels: A survey. *ACM Computing Surveys*, 46(3):42.

Doi, Shinichi and Kazunori Muraki. 1992. Translation ambiguity resolution based on text corpora of source and target language. In *Proc. of COLING*, pages 525–531, Nantes.

Dove, Catherine, Olga Loskutova, and Ruben de la Fuente. 2012. What's your pick: RbMT, SMT or Hybrid? In *Proc. of AMTA*, San Diego.

España-Bonet, Cristina, Jesús Giménez, and Lluís Màrquez. 2010. Discriminative Phrase-Based Models for Arabic Machine Translation. *ACM Transactions on Asian Language Information Processing*, 8:1–20.

España-Bonet, Cristina, Gorka Labaka, Alicia Díaz de Ilarraza, Lluís Màrquez, and Kepa Sarasola. 2011. Hybrid Machine Translation Guided by a Rule-Based System. In *Proc of the MT-Summit*, pages 554–561, Xiamen.

Green, Spence, Sida Wang, Daniel Cer, and Christopher D. Manning. 2013. Fast and Adaptive Online Training of Feature-Rich Translation Models. In *Proc. of the ACL*, pages 311–321, Sofia.

Habash, Nizar and Fatiha Sadat. 2006. Arabic preprocessing schemes for Statistical Machine Translation. In *Proc. of the HLT-NAACL*, pages 49–52, New York.

Heyn, Matthias. 1996. Integrating machine translation into translation memory systems. In *Proc. of EAMT*, pages 113–126, Vienna.

Hutchins, W. John and Harold L. Sommers. 1992. An introduction to Machine Translation. *Academic Press*, 362.

Koehn, Philipp and Barry Haddow. 2012. Towards effective use of training data in Statistical Machine Translation. In *Proc of the WMT*, pages 317–321, Montréal.

Koehn, Philipp, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc of the ACL*, Sapporo.

Liang, Percy, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to Machine Translation. In *Proc. of the ICL-ACL*, pages 761–768, Sydney.

Och, Franz Joseph. 2003. Minimum Error Rate Training In Statistical Machine Translation. In *Proc of the ACL*, pages 160–167, Sapporo.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of Machine Translation. In *Proc. of ACL*, pages 311–318, Philadelphia.

Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal SMT. In *Proc of the ACL*, pages 271–279, Ann Arbor.

Sánchez-Cartagena, Victor, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer Machine Translation rules from scarce bilingual corpora. *Computer Speech and Language. Special Issue on Hybrid MT: Integration of Linguistics and Statistics*, Accepted.

Su, Jinsong, Yang Liu, Yajuan Lü, Haitao Mi, and Qun Liu. 2010. Learning Lexicalized Reordering Models from Reordering Graphs. In *Proc. of the ACL*, pages 12–16, Uppsala.

Tyers, Francis M., Felipe Sánchez-Martínez, and Mikel L. Forcada. 2012. Flexible finite-state lexical selection for rule-based MT. In *Proc of the EAMT*, pages 213–220, Trento.

Wu, Dekai and Pascale Fung. 2009. Semantic roles for SMT: A hybrid two pass model. In *Proc of the NAACL HLT*, pages 13–16, Boulder.

Zollmann, Andreas and Ashish Venugopal.
  2006. Syntax Augmented Machine

Translation via Chart Parsing. In *Proc. of
the WMT*, pages 138–141, New York.