Hybrid machine translation: integration of linguistics and statistics: editorial

José A.R. Fonollosa, *Universitat Politècnica de Catalunya, Jordi Girona 1-3, Barcelona 08034, Spain*

Marta R. Costa-jussà, *Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632, Singapore*

Hybrid systems mix two or more approaches with the aim of combining the advantages of both of them. Hybrid vehicles are a good popular example of the success of this hybrid approach in engineering. Despite their cost, hybrid vehicles are now manufactured and sold with success by many companies and they are a usual choice for certain types of vehicles as taxicabs and delivery trucks.

However, hybrid systems also imply the need of developing, integrating and maintaining two different technologies to attack the same problem. Additionally, a hybrid system may also suffer in some degree from the deficiencies of both approaches.

In Machine Translation (MT), researchers have concentrated most of their efforts during the last years in the development of statistical approaches and tools, as the amount of available data continue to increase. However linguist information and human-derived rules are still useful in many cases, in pre or post-processing the corpora or in the core of the system.

The annual workshops on hybrid approaches to translation (HyTra) are a good example of the interest of the scientific community in the subject and a source of new ideas.

Hybrid approaches are also widely used in commercial MT systems. For companies specialized in MT services, hybrid solutions are the key to personalized high quality systems adapted to the needs of the customers and the available language resources.

This special issue collects a variety of interesting ideas on how to combine linguistic and statistical information in MT. As follows we provide a short overview of all the papers in the volume.

In order to introduce hybrid approaches in the area of MT, this special issue starts with an overview of the *Latest Trends in Hybrid Machine Translation and its Applications* by Costa-jussà and Fonollosa.

The paper *Linguistically-augmented Perplexity-based Data Selection for Language Models* by Toral et al. studies how to reduce the perplexity of the language model including linguistic information in the selection of the training

corpus. The language model itself continues to be an n-gram language model, but the linguistically motivated matching rules are clearly of help in the process of selecting the relevant training sentences. Although the paper does not propose or study any hybrid MT scheme, the hybrid approach for corpus selection in language modeling can be applied in MT. Moreover, it is also a good starting point to develop a similar hybrid selection method of bilingual corpora.

In the paper *A Tree Does Not Make a Well-Formed Sentence: Improving Syntactic String-to-Tree Statistical Machine Translation with More Linguistic Knowledge*, Sennrich et al. describe another example in which additional linguistic information is the key to obtain a competitive translation system. In their paper, a syntactic string-to-tree statistical MT system is clearly improved with the introduction of more morphological constraints and syntactic rules.

The paper *A Generalised Alignment Template Formalism and its Application to the Inference of Shallow-transfer Machine Translation Rules from Scarce Bilingual Corpora* by Sánchez-Cartagena et al. follows a complete different approach. The authors develop here a new method to use scarce parallel corpora to automatically infer a set of shallow- transfer rules. No human experts are required to infer rules, but the automatically learned rules can be post-edited and can co-exist with hand-written rules.

In *Translating Noun Compounds Using Semantic Relations*, Balyan and Chatterjee propose a hybrid approach to solve the problem of translating noun compounds from English to Hindu. The investigated rule-based scheme based on semantic relations is shown to be a good complementary tool for state-of-the art systems as Moses.

In the paper *Translating without In-domain Corpus: Machine Translation Post-Editing with Online Learning Techniques* by Lagarda Arroyo et al., a statistical MT system is proposed as an automatic post-editing module in an online learning framework. The resulting system uses the output of any MT system as input, and it is a good choice for domain adaptation when the documents to be translated have the usual internal (n-gram) repetition property.

The paper *Hybrid Arabic–French Machine Translation using Morphological Processing and Language Analysis* by Mohamed and Sadat presents a rule-based pre-processing for an Arabic–French system based on Moses. The morphological rules reduce the morphology of the source language (Arabic) to a level that makes it closer to that of the target language (French). They also consider the introduction of additional swapping rules for a structural matching between the source language and the target language.

Finally, the paper *Using Decision Tree to Hybrid Morphology Generation of*

Persian Verb for English-Persian Translation by Mahmoudi and Faili proposes hybrid solutions for the translation of verbs from English to morphological rich languages as Persian. The develop approach can be part of a rule-based MT or it could be applied as an independent post-processing step of any statistical or rule-based MT.