Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 32 (2015) 3–10

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Latest trends in hybrid machine translation and its applications☆

Marta R. Costa-jussà[a,*], José A.R. Fonollosa[b]

[a] *Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632, Singapore*
[b] *Universitat Politècnica de Catalunya, Jordi Girona, Barcelona 08034, Spain*

Received 31 October 2014; accepted 1 November 2014
Available online 15 November 2014

**Abstract**

This survey on hybrid machine translation (MT) is motivated by the fact that hybridization techniques have become popular as they attempt to combine the best characteristics of highly advanced pure rule or corpus-based MT approaches. Existing research typically covers either simple or more complex architectures guided by either rule or corpus-based approaches. The goal is to combine the best properties of each type.

This survey provides a detailed overview of the modification of the standard rule-based architecture to include statistical knowledge, the introduction of rules in corpus-based approaches, and the hybridization of approaches within this last single category. The principal aim here is to cover the leading research and progress in this field of MT and in several related applications.


*MSC:* 00-01; 99-00

*Keywords:* Hybridization; Machine translation; Corpus; Rules; Applications

## 1. Introduction

Machine translation (MT) is the area of natural language processing (NLP) that focuses on obtaining a target language text from a source language text by means of automatic techniques. MT is a multidisciplinary field and the challenge has been approached from various points of view including linguistics and statistics. The existence of different perspectives has made possible the proliferation of hybrid methodologies. Hybrid methods focus on combining the best properties of two or more MT approaches. Nowadays, it has become very popular to include rules in statistical MT (SMT) approaches. However, there are also relevant works on enhancing standard rule-based MT (RBMT) by adding statistical knowledge. Recent initiatives such as the three editions of the HyTra workshop[1] show that linguists, engineers and computer scientists actively interact in the interests of building successful hybrid architectures, formulating proposals and conducting experiments.

This survey paper reviews recent methods that combine and hybridize MT approaches in single architectures, and thus, two closely related lines of research fall outside our scope. First, the methodologies of multi-engine combination,

☆ This paper has been recommended for acceptance by Roger K. Moore.
* Corresponding author. Current address: Instituto Politécnico Nacional, Mexico. Tel.: +51 1 5525298370.
[1] http://parles.upf.edu/llocs/plambert/hytra/hytra2014/.

http://dx.doi.org/10.1016/j.csl.2014.11.001

which have been widely studied in MT,[2] as well as in other related areas (e.g. speech recognition). These approaches assemble MT outputs, not MT architectures. And second, the integration of linguistic knowledge into SMT when studies do not consider different MT paradigms. For a survey on this specific topic see Costa-jussà and Farrús (2014).

The rest of the paper is organized as follows. Section 2 explains two classifications of MT approaches. Section 3 reports the main hybridization methods within and across paradigms. Section 4 describes several MT applications with hybrid components. Finally, Section 5 summarizes the main findings of this survey.

## 2. Classification of machine translation

Basically, MT approaches can be classified into different paradigms using two criteria: either *the level of representation* or *the sources of information*.

### 2.1. Level of representation

When classifying MT by level of representation, we can think of the Vauquois pyramid that basically contains: direct, transfer and interlingua approaches.

*Direct*. Approaches at the bottom of the Vauquois pyramid require one single step transformation between source and target, without analysis of the source language and without generation of the target language. Within this category, we might find simple dictionary-based translations.

*Transfer*. Approaches in the middle of the Vauquois pyramid consist of three steps: analysis, transfer and generation. This category includes RBMT, EBMT and SMT approaches.

*Interlingua*. Approaches at the top of the Vauquois pyramid consist of two steps: analysis and generation. The analysis transforms the source language into the interlingua representation and the generation transforms this interlingua representation into the target language. Interlingua is a universal representation of all languages, needing no transfer stage.

Wu (2005) offers observations as to whether a system can be considered direct or transfer depending largely on how much or how little language-specific monolingual analysis is carried out and also how close the intermediate representations are to the source and target texts themselves. Essentially, most of the approaches (other than interlingua) mentioned in this article could be classified as transfer-based engines, with varying degrees of complexity in their transfer, analysis and generation stages.

### 2.2. Sources of information

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.

[2] See references in http://www.statmt.org/survey/Topic/SystemCombination.