Workshop on Hybrid Approaches to Translation: Overview and Developments

Marta R. Costa-jussà, Rafael E. Banchs Institute for Infocomm Research[1]
Reinhard Rapp Aix-Marseille Université, LIF[2]
Patrik Lambert Barcelona Media[3]
Kurt Eberle Lingenio GmbH[4]
Bogdan Babych University of Leeds[5]

[1]{vismrc,rembanchs}@i2r.a-star.edu.sg, [2]reinhardrapp@gmx.de,
[3]patrik.lambert@barcelonamedia.org, [4]k.eberle@lingenio.de,
[5]b.babych@leeds.ac.uk

Abstract

A current increasing trend in machine translation is to combine data-driven and rule-based techniques. Such combinations typically involve the hybridization of different paradigms such as, for instance, the introduction of linguistic knowledge into statistical paradigms, the incorporation of data-driven components into rule-based paradigms, or the pre- and post-processing of either sort of translation system outputs. Aiming at bringing together researchers and practitioners from the different multidisciplinary areas working in these directions, as well as at creating a brainstorming and discussion venue for Hybrid Translation approaches, the HyTra initiative was born. This paper gives an overview of the Second Workshop on Hybrid Approaches to Translation (HyTra 2013) concerning its motivation, contents and outcomes.

1 Introduction

Machine translation (MT) has continuously been evolving from different perspectives. Early systems were basically dictionary-based. These approaches were further developed to more complex systems based on analysis, transfer and generation. The objective was to climb up (and down) in the well-known Vauquois pyramid (see Figure 1) to facilitate the transfer phase or to even minimize the transfer by using an interlingua system. But then, corpus-based approaches irrupted, generating a turning point in the field by putting aside the analysis, generation and transfer phases.

Although there had been such a tendency right from the beginning (Wilks, 1994), in the last years, the corpus-based approaches have reached a point where many researchers assume that relying exclusively on data might have serious limitations. Therefore, research has focused either on syntactical/hierarchical-based methods or on trying to augment the popular phrase-based systems by incorporating linguistic knowledge. In addition, and given the fact that research on rule-based has never stopped, there have been several proposals of hybrid architectures combining both rule-based and data-driven approaches.

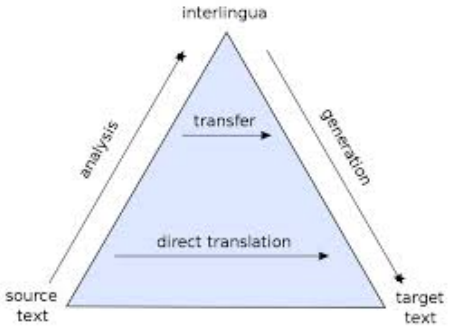

Figure 1: Vauquois pyramid (image from Wikipedia).

In summary, there is currently a clear trend towards hybridization, with researchers adding morphological, syntactic and semantic knowledge to statistical systems, as well as combining data-driven methods with existing rule-based systems.

In this paper we provide a general overview of current approaches to hybrid MT within the context of the Second Workshop on Hybrid Approaches to Translation (HyTra 2013). In our overview, we classify hybrid MT approaches according to the linguistic levels that they address. We then briefly summarize the contributions presented and collected in this volume.

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 1–6,
Sofia, Bulgaria, August 8, 2013.

The paper is organized as follows. First, we motivate and summarize the main aspects of the HyTra initiative. Then, we present a general overview of the accepted papers and discuss them within the context of other state-of-the-art research in the area. Finally, we present our conclusions and discuss our proposed view of future directions for Hybrid MT research.

2 Overview of the HyTra Initiative

The HyTra initiative started in response to the increasing interest in hybrid approaches to machine translation, which is reflected on the substantial amount of work conducted on this topic. Another important motivation was the observation that, up to now, no single paradigm has been able to successfully solve to a satisfactory extent all of the many challenges that the problem of machine translation poses.

The first HyTra workshop took part in conjunction with the EACL 2012 conference (Costa-jussà et al., 2012). The Second HyTra Workshop, which was co-organized by the authors of this paper, has been co-located with the ACL 2013 conference (Costa-jussà et al., 2013). The workshop has been supported by an extensive programme committee comprising members from over 30 organizations and representing more than 20 countries. As the outcome of a comprehensive peer reviewing process, and based on the recommendations of the programme committee, 15 papers were finally selected for either oral or poster presentation at the workshop.

The workshop also had the privilege to be honored by two exceptional keynote speeches:

- *Controlled Ascent: Imbuing Statistical MT with Linguistic Knowledge* by Will Lewis and Chris Quirk (2013), Microsoft research. The intersection of rule-based and statistical approaches in MT is explored, with a particular focus on past and current work done at Microsoft Research. One of their motivations for a hybrid approach is the observation that the times are over when huge improvements in translation quality were possible by simply adding more data to statistical systems. The reason is that most of the readily available parallel data has already been found.

- *How much hybridity do we have?* by Hermann Ney, RWTH Aachen. It is pointed out that after about 25 years the statistical approach to MT has been widely accepted as an alternative to the classical approach with manually designed rules. But in practice most statistical MT systems make use of manually designed rules at least for preprocessing in order to improve MT quality. This is exemplified by looking at the RWTH MT systems.

3 Hybrid Approaches Organized by Linguistic Levels

'Hybridization' of MT can be understood as combination of several MT systems (possibly of very different architecture) where the single systems translate in parallel and compete for the best result (which is chosen by the integrating meta system). The workshop and the papers do not focus on this 'coarse-grained' hybridization (Eisele et al., 2008), but on a more 'fine grained' one where the systems mix information from different levels of linguistic representations (see Figure 2). In the past and mostly in the framework of rule-based machine translation (RBMT) it has been experimented with information from nearly every level including phonetics and phonology for speech recognition and synthesis in speech-to-speech systems (Wahlster, 2000) and including pragmatics for dialog translation (Batliner et al., 2000a; Batliner et al., 2000b) and text coherence phenomena (Le Nagard and Koehn, 2010). With respect to work with emphasis on statistical machine translation (SMT) and derivations of it mainly those information levels have been used that address text in the sense of sets of sentences.

As most of the workshop papers relate to this perspective - i.e. on hybridization which is defined using SMT as backbone, in this introduction we can do with distinguishing between approaches focused on morphology, syntax, and semantics. There are of course approaches which deal with more than one of these levels in an integrated manner, which are commonly refered to as multilevel approaches. As the case of treating syntax and morphology concurrently is especially common, we also consider morpho-syntax as a separate multilevel approach.

3.1 Morphological approaches

The main approaches of statistical MT that exploit morphology can be classified into segmentation, generation, and enriching approaches. The

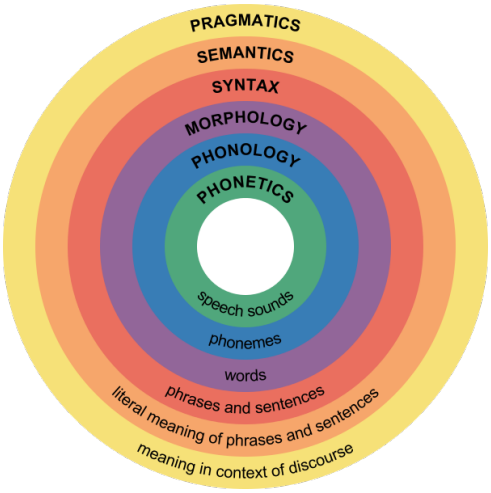

Figure 2: Major linguistic levels (image from Wikipedia).

first one attempts to minimize the vocabulary of highly inflected languages in order to symmetrize the (lexical granularity of the) source and the target language. The second one assumes that, due to data sparseness, not all morphological forms can be learned from parallel corpora and, therefore, proposes techniques to learn new morphological forms. The last one tries to enrich poorly inflected languages to compensate for their lack of morphology. In HyTra 2013, approaches treating morphology were addressed by the following contributions:

- Toral (2013) explores the selection of data to train domain-specific language models (LM) from non-domain specific corpora by means of simplified morphology forms (such as lemmas). The benefit of this technique is tested using automatic metrics in the English-to-Spanish task. Results show an improvement of up to 8.17% of perplexity reduction over the baseline system.

- Rios Gonzalez and Goehring (2013) propose machine learning techniques to decide on the correct form of a verb depending on the context. Basically they use tree-banks to train the classifiers. Results show that they are able to disambiguate up to 89% of the Quechua verbs.

3.2 Syntactic approaches

Syntax had been addressed originally in SMT in the form of so called phrase-based SMT without any reference to linguistic structures; during the last decade (or more) the approach evolved to or, respectively, was complemented by - work on syntax-based models in the linguistic sense of the word. Most such approaches can be classified into three different types of architecture that are defined by the type of syntactic analysis used for the source language and the type of generation aimed at for the target language: tree-to-tree, tree-to-string and string-to-tree. Additionally, there are also the so called hierarchical systems, which combine the phrase-based and syntax-based approaches by using phrases as translation-units and automatically generated context free grammars as rules. Approaches dealing with the syntactic approach in HyTra 2013 include the following papers:

- Green and Zabokrtský (2013) study three different ways to ensemble parsing techniques and provide results in MT. They compute correlations between parsing quality and translation quality, showing that NIST is more correlated than BLEU.

- Han et al. (2013) provide a framework for pre-reordering to make Chinese word order more similar to Japanese. To this purpose, they use unlabelled dependency structures of sentences and POS tags to identify verbal blocks and move them from after-the-object positions (SVO) to before-the-object positions (SOV).

- Nath Patel et al. (2013) also propose a pre-reordering technique, which uses a limited set of rules based on parse-tree modification rules and manual revision. The set of rules is specifically listed in detail.

- Saers et al. (2013) report an unsupervised learning model that induces phrasal ITGs by breaking rules into smaller ones using minimum description length. The resulting translation model provides a basis for generalization to more abstract transduction grammars with informative non-terminals.

3.3 Morphosyntactical approaches

In linguistic theories, morphology and syntax are often considered and represented simultaneously (not only in unification-based approaches) and the same is true for MT systems.

- Laki et al. (2013) combine pre-reordering rules with morphological and factored models for English-to-Turkish.

- Li et al. (2013) propose pre-reordering rules to be used for alignment-based reordering, and corresponding POS-based restructuring of the input. Basically, they focus on taking advantage of the fact that Korean has compound words, which - for the purpose of alignment - are split and reordered similarly to Chinese.

- Turki Khemakhem et al. (2013) present work about an English-Arabic SMT system that uses morphological decomposition and morpho-syntactic annotation of the target language and incorporates the corresponding information in a statistical feature model. Essentially, the statistical feature language model replaces words by feature arrays.

3.4 Semantic approaches

The introduction of semantics in statistical MT has been approached to solve word sense disambiguation challenges covering the area of lexical semantics and, more recently, there have been different techniques using semantic roles covering shallow semantics, as well as the use of distributional semantics for improving translation unit selection. Approaches treating the incorporation of semantics into MT in HyTra 2013 include the following research work:

- Rudnick et al. (2013) present a combination of Maximum Entropy Markov Models and HMM to perform lexical selection in the sense of cross-lingual word sense disambiguation (i.e. by choice from the set of translation alternatives). The system is meant to be integrated into a RBMT system.

- Boujelbane (2013) proposes to build a bilingual lexicon for the Tunisian dialect using modern standard Arabic (MSA). The methodology is based on leveraging the large available annotated MSA resources by exploiting MSA-dialect similarities and addressing the known differences. The author studies morphological, syntactic and lexical differences by exploiting Penn Arabic Treebank, and uses the differences to develop rules and to build dialectal concepts.

- Bouillon et al. (2013) presents two methodologies to correct homophone confusions. The first one is based on hand-coded rules and the second one is based on weighted graphs derived from a pronunciation resource.

3.5 Other multilevel approaches

In a number of linguistic theories information from the morphological, syntactic and semantic level is considered conjointly and merged in corresponding representations (a RBMT example is LFG (Lexical Functional Grammars) analysis and the corresponding XLE translation architecture). In HyTra 2013 there are three approaches dealing with multilevel information:

- Pal et al. (2013) propose a combination of aligners: GIZA++, Berkeley and rule-based for English-Bengali.

- Hsieh et al. (2013) use comparable corpora extracted from Wikipedia to extract parallel fragments for the purpose of extending an English-Bengali training corpus.

- Tambouratzis et al. (2013) describe a hybrid MT architecture that uses very few bilingual corpus and a large monolingual one. The linguistic information is extracted using pattern recognition techniques.

Table 1 summarizes the papers that have been presented in the Second HyTra Workshop. The papers are arranged into the table according to the linguistic level they address.

4 Conclusions and further work

The success of the Second HyTra Workshop confirms that research in hybrid approaches to MT systems is a very active and promising area. The MT community seems to agree that pure data-driven or rule-based paradigms have strong limitations and that hybrid systems are a promising direction to overcome most of these limitations. Considerable progress has been made in this area recently, as demonstrated by consistent improvements for different language pairs and translation tasks.

The research community is working hard, with strong collaborations and with more resources at hand than ever before. However, it is not clear

Morphological	(Toral, 2013)	Hybrid Selection of LM Training Data Using Linguistic Information and Perplexity
	(Gonzales and Goehring, 2013)	Machine Learning disambiguation of Quechua verb morphology
Syntax	(Green and Zabokrtský, 2013)	Improvements to SBMT using Ensemble Dependency Parser
	(Han et al., 2013)	Using unlabeled dependency parsing for pre-reordering for Chinese-to-Japanese SMT
	(Patel et al., 2013)	Reordering rules for English-Hindi SMT
	(Saers et al., 2013)	Unsupervised transduction grammar induction via MDL
Morpho-syntactic	(Laki et al., 2013)	English to Hungarian morpheme-based SMT system with reordering rules
	(Li et al., 2013)	Experiments with POS-based restructuring and alignment based reordering for SMT
	(Khemakhem et al., 2013)	Integrating morpho-syntactic feature for English Arabic SMT
Semantic	(Rudnick and Gasser, 2013)	Lexical Selection for Hybrid MT with Sequence Labeling
	(Boujelbane et al., 2013)	Building bilingual lexicon to create dialect Tunisian corpora and adapt LM
	(Bouillon et al., 2013)	Two approaches to correcting homophone confusions in a hybrid SMT based system
Multilevels	(Pal et al., 2013)	A hybrid Word alignment model for PBSMT
	(Hsieh et al., 2013)	Uses of monolingual in-domain corpora for cross-domain adaptation with hybrid MT approaches
	(Tambouratzis et al., 2013)	Overview of a language-independent hybrid MT methodology

Table 1: HyTra 2013 paper overview.

whether technological breakthroughs as in the past are still possible are still possible, or if MT will be turning into a research field with only incremental advances. The question is: have we reached the point at which only refinements to existing approaches are needed? Or, on the contrary, do we need a new turning point?

Our guess is that, similar to the inflection point giving rise to the statistical MT approach during the last decade of the twentieth century, once again there might occur a new discovery which will revolutionize further the research on MT. We cannot know whether hybrid approaches will be involved; but, in any case, this seems to be a good and smart direction as it is open to the full spectrum of ideas and, thus, it should help to push the field forward.

Acknowledgments

This workshop has been supported by the Seventh Framework Program of the European Commission through the Marie Curie actions HyghTra, IMTraP, AutoWordNet and CrossLingMind and the Spanish "Ministerio de Economía y Competitividad" and the European Regional Development Fund through SpeechTech4all. We would like to thank the funding institution and all people who contributed towards making the workshop a success. For a more comprehensive list of acknowledgments refer to the preface of this volume.

References

Anton Batliner, J. Buckow, Heinrich Niemann, Elmar Nöth, and Volker Warnke, 2000a. *The Prosody Module*, pages 106–121. New York, Berlin.

Anton Batliner, Richard Huber, Heinrich Niemann, Elmar Nöth, Jörg Spilker, and K. Fischer, 2000b. *The Recognition of Emotion*, pages 122–130. New York, Berlin.

Pierrette Bouillon, Johanna Gerlach, Ulrich Germann, Barry Haddow, and Manny Rayner. 2013. Two approaches to correcting homophone confusions in a hybrid machine translation system. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Rahma Boujelbane, Mariem Ellouze khemekhem, Siwar BenAyed, and Lamia HadrichBelguith. 2013. Building bilingual lexicon to create dialect tunisian corpora and adapt language model. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Marta R. Costa-jussà, Patrik Lambert, Rafael E. Banchs, Reinhard Rapp, and Bogdan Babych, editors. 2012. *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*. Association for Computational Linguistics, Avignon, France, April.

Marta R. Costa-jussà, Patrik Lambert, Rafael E. Banchs, Reinhard Rapp, Bogdan Babych, and Kurl Eberle, editors. 2013. *Proceedings of the Second Workshop on Hybrid Approaches to Translation (HyTra)*. Association for Computational Linguistics, Sofia, Bulgaria, August.

Andreas Eisele, Christian Federmann, Hans Uszkoreit, Hervé Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann, and Yu Chen. 2008. Hybrid machine translation architectures within and beyond the euromatrix project. In John Hutchins and Walther v.Hahn, editors, *12th annual conference of the European Association for Machine Translation (EAMT)*, pages 27–34, Hamburg, Germany.

Annette Rios Gonzales and Anne Goehring. 2013. Machine learning disambiguation of quechua verb morphology. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Nathan Green and Zdenek Zabokrtský. 2013. Improvements to syntax-based machine translation using ensemble dependency parsers. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Dan Han, Pascual Martinez-Gomez, Yusuke Miyao, Katsuhito Sudoh, and Masaaki NAGATA. 2013. Using unlabeled dependency parsing for pre-reordering for chinese-to-japanese statistical machine translation. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

An-Chang Hsieh, Hen-Hsen Huang, and Hsin-Hsi Chen. 2013. Uses of monolingual in-domain corpora for cross-domain adaptation with hybrid mt approaches. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Ines Turki Khemakhem, Salma Jamoussi, and Abdelmajid Ben Hamadou. 2013. Integrating morpho-syntactic feature in english-arabic statistical machine translation. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

László Laki, Attila Novak, and Borbála Siklósi. 2013. English to hungarian morpheme-based statistical machine translation system with reordering rules. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden, July. Association for Computational Linguistics.

Will Lewis and Chris Quirk. 2013. Controlled ascent: Imbuing statistical mt with linguistic knowledge. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Shuo Li, Derek F. Wong, and Lidia S. Chao. 2013. Experiments with pos-based restructuring and alignment-based reordering for statistical machine translation. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Santanu Pal, Sudip Naskar, and Sivaji Bandyopadhyay. 2013. A hybrid word alignment model for phrase-based statistical machine translation. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Raj Nath Patel, Rohit Gupta, Prakash B. Pimpale, and Sasikumar M. 2013. Reordering rules for english-hindi smt. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Alex Rudnick and Michael Gasser. 2013. Lexical selection for hybrid mt with sequence labeling. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Markus Saers, Karteek Addanki, and Dekai Wu. 2013. Unsupervised transduction grammar induction via minimum description length. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

George Tambouratzis, Sokratis Sofianopoulos, and Marina Vassiliou. 2013. Language-independent hybrid mt with presemt. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Antonio Toral. 2013. Hybrid selection of language model training data using linguistic information and perplexity. In *ACL Workshop on Hybrid Machine Approaches to Translation*, Sofia.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin, Heidelberg, New York.

Yorick Wilks. 1994. Stone soup and the french room: The empiricist-rationalist debate about machine translation. *Current Issues in Computational Linguistics: in honor of Don Walker*, pages 585–594. Pisa, Italy: Giardini / Dordrecht, The Netherlands: Kluwer Academic.