HyTra 5 Proceedings Preview

Introductory Remarks

This volume contains the papers presented at HyTra-5: *HyTra-5: Fifth Workshop on Hybrid Approaches to Translation* held on June 1, 2016 in Riga, including the invited talk entitled *Hybrid MT in Academia & Industry* held by Prof. Andy Way from the Adapt Centre Dublin and the School of Computing of Dublin City University (DCU) and some statements and results from the panel session of the workshop that focused on questions about the best way of cooperation between academic research and industrial development, hybridization and integration of company translation data into customized MT systems.
Besides Andy Way, participants of the panel session were Alexander Fraser (from Center for Information and Language Processing (CIS) of Ludwig-Maximilians-Universität Munich), Maja Popović (from Humboldt University of Berlin), Tony O'Dowd (from KantanMT) and Bogdan Babych (from the Center of Translation Studies of the University of Leeds) as representative of the organizers of the workshop. Besides invited talk and panel session there were 3 accepted papers presented in the workshop.

Workshop Topic and Content

The intent of the Fifth Workshop on Hybrid Approaches to Translation (HyTra-5) was to continue developing and empowering the research agenda in the area of Hybrid MT and to follow up the discussions and results of the previous workshops of the series:

The first edition of HyTra was held (together with the ESIRMT workshop) as a joint 2-day workshop at EACL 2012, in Avignon, France. The second, third and fourth editions of HyTra were held as 1-day workshops at ACL 2013 (Sofia, Bulgaria), EACL 2014 (Gothenburg, Sweden) and ACL 2015 (Beijing, China) respectively.

These first four editions of HyTra brought together a good number of researchers, students and industrial practitioners working on diverse problems and applications related to hybrid machine translation. We are happy that in all workshop of the series the participants could benefit from the inspiring multicultural and multi-disciplinary atmosphere, the ambitious and focused technical program and the renowned keynote speakers: Christof Monz and Philipp Koehn (HyTra-1), Hermann Ney, William Lewis and Chris Quirk (HyTra-2), Hans Uszkoreit and Joakim Nivre (HyTra-3) and Hinrich Schutze and Gerard de Melo (HyTra-4).
In all workshops full paper and poster sessions have allowed for fruitful, creative and interdisciplinary discussions, comparing and contrasting diverse ways to integrate different MT paradigms to improve the state-of-the-art in Machine Translation resulting in proceedings that put together high-quality papers experimenting with various research topics including statistical approaches integrating morphological, syntactic, semantic and rule-based information. In addition to the proceedings of the workshop, one contributed volume, composed of papers selected from the first three editions of HyTra, is on the way and will be soon published by Springer. Workshop programs in HyTra-3 and HyTra-4 have also been enriched with industrial sessions, in which current MT problems and

approaches have been surfaced and discussed by industry players from both the technical and the commercial perspectives.
It is in the spirit of these four previous editions that the fifth workshop has been organized and held in Riga, the main objective still being to motivate the cooperation and interaction between the different human components of such a highly interdisciplinary and multidisciplinary field - as translation is - namely translators, engineers, computer scientists, mathematicians and linguists, as well as to foster innovation and creativity in the Hybrid Machine Translation research community.

Given the complementarity and, consequently, mutual attractiveness of data-driven and rule-based MT, it is natural that the boundaries among the 'philosophies' have narrowed significantly nowadays and that the interest in hybridization and system combination has constantly increased. So the question meanwhile is not so much whether a system should be 'pure' SMT or RBMT, rather it is what the combined architecture or architectures should look like in detail and whether there are new methods that could be integrated. As a starting point of such considerations in general, and of the presentations and discussions in our fifth HyTra workshop in particular, that you may find in condensed form in these proceedings, the following types of hybrid MT may be identified as a kind of today's standard of hybrid MT:

(1) SMT models augmented with morphological, syntactic or semantic information;
(2) Rule-based MT systems using parallel and comparable corpora to improve results by enriching their lexicons and grammars and by applying new methods for disambiguation;
(3) MT system output combination based on different MT paradigms (including voting systems);
(4) various post-editing approaches.

As said, the question is how to develop, mix and extend these lines further.

Acknowledgement

We are grateful to EAMT and its organizing committee to have supported HyTra-5. We are also grateful to the members of our program committee for reviewing the papers and, needless to say, we are especially grateful to our panelists and Andy Way, who in addition provided us with a very interesting and deep invited talk.

Program

The program consisted of the invited talk, the panel and the presentation of the accepted papers (cf. http://glicom.upf.edu/hytra2016/program.html). We follow this order below. Additional information about the workshop, its organization and the program can be found on the HyTra5 website http://glicom.upf.edu/hytra2016/program.html.

Invited Talk

The invited talk was held by Prof. Andy Way from the Adapt Centre Dublin and the School of Computing of Dublin City University (DCU). Please find below an abstract and the slides of the talk.

Abstract

Slides

Panel Session

The panel discussion focused on questions about the best way of cooperation between academic research and industrial development, hybridization and integration of company translation data into customized MT systems. In particular, the questions addressed to the panelists were the following:

1. What is the best way of cooperation between academic research and industrial development?
   - Concerning maturity of systems: Should academic research partake in developing core engines for industrial application which are optimal with respect to coverage or should it instead better concentrate on new algorithms and create only prototypes and let industry fill out gaps and add features of robustness?
   - If it is prototypes, can the value of the approach be estimated correctly without filling these gaps before?
2. How do you see the role of hybridization: is it really needed?
   - If so, are there differences depending on the specific application in real life and the needs of customers?
3. Normally, companies already possess information in the field of translation (translation skills and resources) when integrating an MT solution. Which MT architectures are particularly suited for integrating such resources and skills ?
4. Do we need a new shared task for optimizing research in the field of industrial-academic cooperation ?
5. Are there other success stories of academia/industry cooperation?
6. What are the flows of data from industry to academia (how can we get this working)?
7. How to organize increasing cooperation between academia and industry with respect to training people?

The participants of the panel were:

Alexander Fraser	Center for Information and Language Processing (CIS) of Ludwig-Maximilians-Universität Munich
Tony O'Dowd	KantanMT
Maja Popović	Humboldt University of Berlin
Andy Way	Adapt Centre Dublin and School of Computing of Dublin City University (DCU)
Bogdan Babych	Center of Translation Studies of the University of Leeds (for the organizers)

Some results from the panel and statements are the following:

Maja Popović argued with respect to question 2 that hybridization is not an end in itself, but has to be assessed from the perspective of the purpose of a corresponding system in question and from the respective potiental gain in quality that may be reached by the hybridization investigated, i.e. given a specific context '*whatever might deliver suitable results should be tried*.'
Concerning question 4 she expected a new shared task to be useful, provided it satisfies some conditions, which are the following:
- the goals of the shared task should be defined by industry;
- the task consists of developing systems which concentrate on these goals,

- the evaluation identifies the system which performs best in the area defined by the goals;
- this includes investigation and development of evaluation techniques suitable for industry goals.

Summarizing, Maja Popović emphasized that the whole process of working together in a shared task provides a good opportunity for both "sides" (industry and research) to learn more about each other and to get a better understanding of each other's perspective.

With respect to questions 1, 5, 6 and 7 **Andy Way** drew the attention to an inherent conflict concerning the distribution of work between academia and industry *'namely the fact that just as the technology has proven to be valuable in various industry workflows, industry is bemoaning the fact that there isn't enough suitably qualified staff coming from universities'*. He continues by a conclusion he draws from this observation: *'A point of view that I would take is that industry is partially responsible for this state of affairs, as 'top' MT academics are (quite understandably) being hoovered up by large organizations, with the consequence that the few MT Centres of Excellence that exist become even fewer*' (and can educate and train only fewer students – too few to satisfy the needs of industry). His summarizing conclusion: *'Industry can't have it both ways: if they recruit the leaders of large, renowned academic groups, who were training the new MT developers of tomorrow, they shouldn't be surprised when the number of such potential recruits falls away ...* '

We can take from this that communication between industry and academia should be much deeper and should discuss not only short termed goals from a scientific and engineering viewpoint, but should include considerations about the development as a whole in the medium and long run, including questions of education, career and personnel development

Bogdan Babych took from the panel discussion and the workshop as a whole the following summarizing statement about the state and role of hybrid machine translation:

The concept of a hybrid MT becomes less clear nowadays, as both SMT and RBMT paradigms increasingly share ideas and technologies for resource development and runtime engine operation. This has lead to suggestions that the hybrid MT is no longer a useful concept, because every modern MT approach is in a certain respect “hybrid”. There is a pressure within the current research paradigm in MT to avoid this concept in favour of the approach-specific terms such as factored MT, syntax-based MT, semantic MT, etc. (A similar terminological pressure earlier merged Example-based MT (EBMT) the core SMT approaches – as has been discussed by Andy Way in his keynote speech). On the one hand, core SMT techniques increasingly integrate morphological and syntactic models; the neural MT captures long-distance translation equivalent dependencies beyond N-gram level. On the other hand, RBMT systems, which traditionally had a much longer development cycle, increasingly use statistical models for speeding up the development of morphological, syntactic and semantic resources for the analysis, generation and transfer stages, and for the runtime operation of modules in guiding disambiguation and translation equivalent choices. Many RBMT systems use statistical techniques as standard, which again adds pressure on redefining the scope of the term Hybrid MT.

For the Hybrid MT to continue as a viable research paradigm there is a need to communicate to the research community its renewed sense of purpose, in light of recent technological developments in the field. There are several ways, in which this can be achieved:

1. There is a suggestion (Antonio Toral) to redefine hybrid MT to include approaches which combine Neural MT with SMT + RBMT in different combinations, as Neural MT approaches currently do not 'talk' to SMT and RBMT paradigms yet, in a way that RBMT and SMT started to develop common frameworks.
2. Hybrid MT can be redefined as an approach where hybrid techniques are used at the core in a systematic way, not just added as additional flavour features to the core SMT or RBMT engines (the closest to this is Systran's approach, although it may not be there yet). The defining question for narrowing down true hybrid approaches in this way would be – to understand and systematically cover the phenomena which cannot be treated within the 'hybrid extensions' paradigm: e.g., parts of the morphology which are not covered by the factor-based SMT models, syntax (e.g., linguistic constructions) which go beyond the capacity of the syntax-based SMT models, etc.)

In my view, is important to keep a clear distinction between RBMT, SMT and hybrid approaches along the following lines:

RBMT systems, as compared to SMT and Neural MT, present information about language in a more compact and generalised way; the fact that they are more robust against changes in the domain on which resources are trained points out to the important feature: RBMT aims at going beyond 'seen' training examples; or more broadly – beyond Tarsky's definition of meaning, aims at capturing translation strategies and generating translation equivalents along those strategies, without the need for them to be present in the training data. RBMT effectively aims at reducing the entropy of the language description (where entropy is understood as a description length of the phenomenon). Reducing entropy allows RBMT systems to generally have a smaller footprint and to operate with observable representations, which can be corrected and individually checked for their effect on the translation performance.

SMT and Neural MT systems effectively attempt to account for the same phenomena with much larger datasets, keeping the entropy of the training data within their larger representations, achieving better performance at the cost of giving up on generalization. Currently, they make little attempt or have task to reduce the description length or arrive at observable and correctable representations.

From this perspective Hybrid MT can be redefined as an attempt to understand the mechanisms of arriving at compact linguistic generalisations and interpretable processing models using of large statistically processed datasets of parallel, comparable and monolingual corpora. It may not necessarily lead to improvement in MT performance, as this is initially an attempt to properly do housekeeping within MT architecture and the value is internal for MT research. However, if linguistically meaningful representations and generalisation strategies can be crystallised from statistically analysed corpora, without the loss of coverage or performance, the added benefits would be:

- MT systems with smaller footprint covering the same phenomena which can be made available on a larger range of platforms
- Extended and more systematic coverage of unseen translation equivalents derived via generalizations (extending the coverage for less frequent phenomena and improving quality for under-resourced languages)
- Greater stability across subject domains, in cases of the domain mismatch between the training data and a given translation project

These tangible improvements go beyond the internal research value of the hybrid approaches, but their benefits may not necessarily be visible in the mainstream evaluation framework on large standard datasets – as they will address specific scenarios of building and using hybrid MT systems, e.g., the scenario of under-resourced languages, or coverage of out-of-domain or rare language data.

There is an argument (given by Jörg Tiedemann during the panel discussion) that from a cognitive perspective, the language and translation mechanisms should not require generalizations: they may work on the large-dataset basis, without explicitly generalising – by keeping all the available data ready for the processing stage. However, in my view that even if linguistic generalizations, such as grammar writing, can be epiphenomenal, there should be some added value in reducing the description length of the translation tasks. Cognitive plausibility may be justified here not in a way of explicitly writing grammars, but via appealing to human memory processing, which may involve loosing individual examples, but reusing general principles for dealing with unseen cases.

A redefined understanding of Hybrid MT, therefore, could be that this approach takes the issue of the description length of the system more seriously, and aims at preserving the performance of large-scale SMT and Neural MT systems using more compact, more reusable and more general data representations and processing models, effectively keeping all the necessary data and performance, but bringing it to a higher reusability level.

Papers