Tutorials

September 8th, 2011

Tutorial 1 (T1)

Speaker:

Prof. Dekai Wu (The Hong Kong University of Science and Technology, China)

Prof. Wu is a pioneer in the fields of syntactic SMT and semantic SMT, and built the world's first large-scale commercially deployed web translation system in 1995. He received his PhD in Computer Science from the University of California at Berkeley, and was a postdoctoral fellow at the University of Toronto (Ontario, Canada) prior to joining HKUST in its founding year in 1992. He received a BS in Computer Engineering from the University of California at San Diego (Revelle College departmental award, cum laude, Phi Beta Kappa) in 1984, and an Executive MBA from Kellogg and HKUST in 2002. He has been a visiting researcher at Columbia University in 1995-96, Bell Laboratories in 1995, and the Technische Universität München (Munich, Germany) during 1986-87. Prof. Wu serves as Associate Editor of *AI Journal* and on the Editorial Boards of *Machine Translation* and *Journal of Natural Language Engineering*. He has co-chaired the annual SSST (Syntax, Semantics and Structure in Statistical Translation) workshops since 2006, and has also served as Co-Chair for EMNLP-2004, and on the Editorial Board of *Computational Linguistics* and as Associate Editor of *ACM Transactions on Speech and Language Processing*, the Organizing Committee of ACL-2000 and WVLC-5 (SIGDAT 1997), and the Executive Committee of the Association for Computational Linguistics (ACL).

Title:

Syntactic SMT and Semantic SMT

Abstract:

Over the past twenty years, we have attacked the historical methodological barriers between statistical machine translation and traditional models of syntax, semantics, and structure. In this tutorial, we will survey some of the central issues and techniques from each of these aspects, with an emphasis on `deeply theoretically integrated' models, rather than hybrid approaches such as superficial statistical aggregation or system combination of outputs produced by traditional symbolic components. On syntactic SMT, we will explore the trade-offs for SMT between learnability and representational expressiveness. After establishing a foundation in the theory and practice of stochastic transduction grammars, we will examine very recent new approaches to automatic unsupervised induction of various classes of

transduction grammars. We will show why stochastic linear transduction grammars (LTGs and LITGs) and their preterminalized variants (PLITGs) are proving to be particularly intriguing models for the bootstrapping of inducing full-fledged stochastic inversion transduction grammars (ITGs). On semantic SMT, we will explore the trade-offs for SMT involved in applying various lexical semantics models. We will first examine word sense disambiguation, and discuss why traditional WSD models that are not deeply integrated within the SMT model tend, surprisingly, to fail. In contrast, we will show how a deeply embedded phrase sense disambiguation (PSD) approach succeeds where traditional WSD does not. We will then turn to semantic role labeling, and discuss the challenges of early approaches of applying SRL models to SMT. Finally, on semantic MT evaluation, we will explore some very new human and semi-automatic metrics based on semantic frame agreement. We show that by keeping the metrics deeply grounded within the theoretical framework of semantic frames, the new HMEANT and MEANT metrics can significantly outperform even the state-of-the-art expensive HTER and TER metrics, while at the same time maintaining the desirable characteristics of simplicity, inexpensiveness, and representational transparency.

Date and time: September 19th, 09:00-12:00 <CHANGED>

Tutorial 2 (T2)

Speakers:

Dr. Yanjun Ma (Baidu, China)

Yanjun Ma is a researcher of machine translation and natural language processing at Baidu. He holds a PhD of Computer Science from Dublin City University. Yanjun's research covers a number of topics in machine translation, including statistical alignment models, hybrid machine translation, machine translation confidence estimation, domain adaptation etc. As a member of Baidu online translation team, he also conducts research on various application scenarios of online translation services. Yanjun has authored and co-authored more than 20 research publications, and serves as a regular PC member/reviewer for most major conferences and journals in areas of machine translation and natural language processing.

Dr.Yifan He (Dublin City Univesity, Ireland)

Yifan He recently passed his PhD defense at Centre for Next Generation Localisation (CNGL), School of Computing, Dublin City University. His PhD research focuses on MT quality estimation methods, especially the methods that help the integration of MT into TM systems, and has led to a pending US patent and a number of research publications at ACL and COLING.

Prof.Josef van Genabith (Dublin City Univesity, Ireland)

Josef van Genabith is a professor in the School of Computing, Dublin City University, Ireland, and the director of the Centre for Next Generation Localisation, a large industry-academia partnership across four universities and nine industry partners.

Title:

From the Confidence Estimation of Machine Translation to the Integration of MT and Translation Memory

Abstract:

In this tutorial, we cover techniques that facilitate the integration of Machine Translation (MT) and Translation Memory (TM), which can help the adoption of MT technology in localisation industry. The tutorial covers four parts: i) brief introduction of MT and TM systems, ii) MT confidence estimation measures tailored for the TM environment, iii) segment-level MT and MT integration, iv) sub-segment level MT and TM integration, and v) human evaluation of MT and TM integration.

We will first briefly describe and compare how translations are generated in MT and TM systems, and suggest possible avenues to combines these two systems. We will also cover current quality / cost estimation measures applied in MT and TM systems, such as the fuzzy-match score in the TM, and the evaluation/confidence metrics used to judge MT outputs.

We then move on to introduce the recent developments in the field of MT confidence estimation tailored towards predicting post-editing efforts. We will especially focus on the confidence metrics proposed by Specia et al., which is shown to have high correlation with human preference, as well as post-editing time.

For segment-level MT and TM integration, we present translation recommendation and translation re-ranking models, where the integration happens at the 1-best or the N-best level, respectively. Given an input to be translated, MT-TM recommendation compares the output from the MT and the TM systems, and presents the better one to the post-editor. MT-TM re-ranking, on the other hand, combines k-best lists from both systems, and generates a new list according to estimated post-editing effort. We observe high precision of these models in automatic and human evaluations, indicating that they can be integrated into TM environments without the risk of deteriorating the quality of the post-editing candidate.

For sub-segment level MT and TM integration, we try to reuse high quality TM chunks to improve the quality of MT systems. We can also predict whether phrase pairs derived from fuzzy matches should be used to constrain the translation of an input segment. Using a series of linguistically- motivated features, our constraints lead both to more

consistent translation output, and to improved translation quality, as is measured by automatic evaluation scores.

Finally, we present several methodologies that can be used to track post-editing effort, perform human evaluation of MT-TM integration, or help translators to access MT outputs in a TM environment.

Date and time: September 19th, 13:30-16:30

Tutorial 3 (T3)

Speaker:

Prof. Alon Lavie (Carnegie Mellon University and Association for Machine Translation in the Americas (AMTA), USA)

Dr. Alon Lavie is an Associate Research Professor at the Language Technologies Institute at Carnegie Mellon University, where he directs a research group in the area of Machine Translation (MT). His current main research projects focus on the design of syntax-based data-driven approaches to Machine Translation, multi-engine Machine Translation system combination, and MT evaluation. Dr. Lavie is also the co-founder and President of Safaba Translation Solutions - a CMU spin-off company that develops Machine Translation solutions for commercial enterprises and Language Service Providers. Dr. Lavie is currently serving as President of the Association for Machine Translation in the Americas (AMTA).

Title:

Evaluating the Output of Machine Translation Systems

Abstract :

This half-day tutorial provides a broad overview of how to evaluate translations that are produced by machine translation systems. The range of issues covered includes a broad survey of both human evaluation measures and commonly-used automated metrics, and a review of how these are used for various types of evaluation tasks, such as assessing the translation quality of MT-translated sentences, comparing the performance of alternative MT systems, or measuring the productivity gains of incorporating MT into translation workflows.

Date and Time: September 19th, 13:30-16:30

Tutorial 4 (T4)

Speakers :

Mr. Mirko Plitt (Autodesk, Switzerland)

Mirko Plitt is the Senior Manager, Language Technologies, in the Localization department of Autodesk, the leading maker of CAD software best known for AutoCAD. In 2001, Mr. Plitt led one of the industry's first implementations of on-the-fly machine translation, to translate Autodesk's product support articles. He later oversaw the introduction of a company-wide authoring and translation management system used by hundreds of technical writers, localization engineers and translators. He then was responsible for the integration of machine translation into Autodesk's translation ecosystem, arguably the industry's largest Moses production deployment. Most recently, Mr. Plitt organized the First Swiss Translation Unconference.

Dr. Ventsislav Zhechev (Autodesk, Switzerland)

Dr. Ventsislav Zhechev is a Computational Linguist at Autodesk in Switzerland, maintaining their in-house Moses-based MT systems. Previously, he was a post-doctoral researcher affiliated to the CNGL in DCU, Dublin, Ireland, working on the EuroMatixPlus project. He is the organiser of the annual Joint EM+/CNGL Workshop and of the Fourth MT Marathon, which took place in DCU 25–30 January 2010. His current research focuses on the integration of Machine Translation and Translation Memory technologies. Previously he developed and released a popular open-source tool for the automatic generation of parallel treebanks.

Title:

Productive Use of MT in Localization

Abstract :

Localization is a term mainly used in the software industry to designate the adaptation of products to meet local market needs. At the center of this process lies the translation of the most visible part of the product – the user interface – and the product documentation. Not surprisingly, the localization industry has therefore long been an extensive consumer of translation technology and a key contributor to its progress.

Software products are typically released in recurrent cycles, with large amounts of content remaining unchanged or undergoing only minor modifications from one release to the next. In addition, software development cycles are short, forcing translation to start while the product is still undergoing changes, so that localized products can reach global markets in a timely fashion. These two aspects result in a heavy dependency on the efficient handling of translation updates. It is only natural that the software industry turned to software-based productivity tools to automate the recycling of translations (through translation memories) and to support the management of the translation workflow (through translation management systems).

Machine translation is a relatively recent addition to the localization technology mix, and not yet as widely adopted as one would expect. Its initial use in the software industry was for more accessory content which is otherwise often left untranslated, e.g. product support articles and antivirus alerts with their short lifecycle. The expectation had however always been that MT could one day be deployed on the bulk of user interface and product documentation, due to the expected process efficiencies and cost savings. While MT is generally still not considered "good" enough to be used raw on this type of content, it has now become an integral part of translation productivity environments, thereby transforming translators into post-editors.

The tutorial will provide an overview of current localization practices and challenges, with a special focus on the role of translation memory and translation management technologies. As a use case of the integration of MT in such an environment, we will then present the approach taken by Autodesk with its large set of Moses engines trained on custom data. Finally, we will explore typical scenarios in which machine translation is employed in the localization industry, using practical examples and data gathered in different productivity and usability tests.

Date and time: September 19th, 09:00-12:00 <CHANGED>