Bologna Translation Service: Online translation of course syllabi and study programmes in English

Heidi Depraetere
CrossLang
W. Wilsonplein 7
9000 Gent, Belgium
heidi@crosslang.com

Joachim Van den Bogaert
CrossLang
W. Wilsonplein 7
9000 Gent, Belgium
joachim@crosslang.com

Joeri Van de Walle
CrossLang
W. Wilsonplein 7
9000 Gent, Belgium
joeri@crosslang.com

Abstract

BTS – Bologna Translation Service – is an ICT PSP EU-funded project which specializes in the automatic translation of study programmes. At the core of the BTS framework are several machine translation (MT) engines through which web-based translation services are offered. The fully integrated BTS architecture includes a translation engine coupling rule-based and statistical MT with automatic post-editing (APE). In this paper, we first describe the motivating factors behind the provision of such a service. Following this, we give an overview of the BTS framework as a whole, with particular emphasis on the MT components, and provide a real world use case scenario in which BTS MT services are exploited.

1 Introduction

There is a continuing increasing need for educational institutes to provide course syllabi documentation and other educational information in English. Access to translated course syllabi and degree programmes plays a crucial role in the degree to which universities effectively attract foreign students and, more importantly, has an impact on international profiling. To present all education information in English is a major challenge for most higher education institutes. The figures and trends show that investment in traditional human translation services is prohibitive, consequently course materials and degree programmes are often provided in the local language only.

The Bologna Translation Service aims at providing a solution to this problem by offering a low-cost, web-based, high-quality machine translation (MT) service. The project will make use of existing rule-based and statistical machine translation technologies and tailor them in order to try and produce the best possible quality for the syllabi translations. The BTS project will demonstrate the customization, integration and validation of software components and data, and will showcase high-quality MT output for citizens, institutions and businesses, to avail of university programmes of study they are currently unaware of.

2 BTS Overview

BTS is a 24 month project (beginning March 2011) funded by the EU under the Information and Communication Technologies Policy Support Programme (ICT PSP). ICT PSP aims at stimulating innovation and competitiveness through the wider uptake and best use of ICT by citizens, governments and businesses. In the sections following we will describe the overall context and implementation strategy of the BTS project.

2.1 The BTS Consortium

BTS comprises a dynamic industry-academia consortium, each member of which brings significant experience and expertise to some facet of the service.

Cross Language brings to the table a team of professionals with expertise in process engineering, computational linguistics, knowledge management, business automation and all aspects of the language market. Convertus AB contributes its Convertus Syllabus Translator (CST) solution, a hybrid machine translation technology that was developed specifically for translating Swedish course materials into English and vice versa. As a technology-driven language service provider,

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 29–34
Leuven, Belgium, May 2011

Traslán contributes their in-house developed Machine Translation technology. Eleka has a long history in developments in the area of language engineering: text and grammar checkers, search engines, translation memories, machine translation, information retrieval and speech technologies. They were the leader of the Opentrad consortium (www.opentrad.com) that developed both Apertium (Ramírez-Sánchez, Sánchez-Martínez, Ortiz-Rojas, Pérez-Ortiz, & Forcada, 2006) and Matxin (Alegria, et al., 2005) open source MT engines (http://matxin.souceforge.net). The Koç University Natural Language Processing (KUNLP) Group, finally, is actively involved in research in machine translation, statistical language modelling, word sense disambiguation, morphology, and parsing.

2.2 BTS Objectives

Bologna will provide a web-based, high-quality, user-oriented, easily accessible, low cost machine-translation service for the educational domain, the Bologna Translation Service (BTS). The service will provide translation of syllabi and study programmes from seven local languages (Finnish, Dutch, French, Spanish, Portuguese, German and Turkish) to English, and from English to Chinese.

The high translation quality of the Bologna project will be achieved through the use of front-edge technological achievements in machine translation (a combination of statistical and rule-based methods), domain-adaptation of the translation engines, the use of automatic post-editing facilities, manual editing, and a memory function. The automatic post-editing facilities and the memory function will increase the translation quality gradually as the service is being used, and the need for manual post-editing will decrease. Consistent use of English terminology within each language pair will be ensured.

The main criteria for success will be:

- user satisfaction as reported by the Bologna user group
- the number of active users, as shown by logs
- the improvement in quality, measured on the basis of standard evaluation methods

3 Data Collection

Quite a substantial part of the project will be devoted to data collection for training the statistical machine translation engines.

Freely available, already aligned corpora of high quality (e.g. the Europarl corpus (Koehn, 2005) and corpora from the EuroMatrix plus project, JRC-ACQUIS Multilingual Parallel Corpus (Steinberger, et al., 2006), OPUS corpus (Tiedemann & Nygaard, 2004), DPC (Macken, Trushkina, & Rura, 2007) etc.) will be used as a basis. This data will be extended with the data sets provided by the University Partners.

The freely available, although not specifically in-domain, corpora will provide a large amount of general domain training data which, after due domain adaptation, can be used to augment any existing in-domain data. Domain adaptation is an active area (see ACL 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP)) with several mature techniques that have been demonstrated to perform well on machine translation. Successful techniques include using monolingual texts to generate bilingual parallel data for training (Bertoldi & Marcello, 2009), language model adaptation (Eck, Vogel, & Waibel, 2004), and translation model adaptation (Nguyen, et al., 2009). In addition, web-crawling techniques will be used. We are confident that we will collect 500K sentence pairs of in-domain data for English into Chinese, Dutch, French, German, Spanish and Portuguese into English while probably less for Finnish and Turkish into English.

4 Machine Translation in BTS

BTS will use pre-existing MT frameworks, solutions and methods to achieve its goals. The main challenges are adapting these solutions for the language pairs described in section 2.2, and compiling all available technology into a user-friendly, high-quality translation system. The following sections describe the systems and strategies that will be used.

4.1 Convertus Syllabus Translator and Convertus Hybrid Translator

The Convertus Syllabus Translator (CST) modules will already be available at the start of the project, including the pre- and post-processing modules which will be ported to the Bologna project languages, the CST post-editing user interface and the CST system integration software. The development of any additional MT system will be fit into this modular framework. Parsers automatically trained in the open-source MaltParser (Nivre, Hall, & Nilsson, 2006) framework (http://maltparser.org) will be used. For the pre-

processing modules of the CST framework, Koç University will supply their Turkish morphological analyser (Sak, Güngör, & Saraçlar, 2010), (Yuret & Ture, 2006), (Oflazer, 1993), and the open-source Finish morphology tool Omorfi (Pirinen, 2008) can be used from the start of the project. The Convertus Hybrid Translator service includes spell-checking, translation, and an interface for manual post-editing. The interface can also be optionally connected to a translation memory that grows as the service is being used, implying in which case that translation quality is also increased. Translation proceeds sentence by sentence and the first step is a search of the source-side of the TM in memory. When no hit is found (above some pre-defined fuzzy match threshold), translation proper is activated. By means of different symbols, the interface distinguishes the translation segments retrieved from memory from those that are newly produced by the MT engine. Edits Changes made during an editing session are also marked by a specific symbol. This marking has been found to be very helpful by the users. Automatic post-editing is provided as a special service, improving translation quality still further and reducing the manual post-editing effort.

Open-source taggers, i.e. HunPos (Halácsy, Kornai, & Oravecz, 2007) and MXPOST (Ratnaparkhi, 1996) will be used for tagging English and Chinese, respectively, as a basis for grammar-based automatic post-editing.

For training and tuning the SMT systems, the freely available Moses (Koehn, et al., 2007) software tools and relevant wrapper scripts included in OpenMaTrEx (Dandapat, Forcada, Groves, Penkale, Tinsley, & Way, 2010) will be used.

4.2 Apertium

The Apertium open-source MT platform will be available for building rule-based MT systems for Spanish and Portuguese. Apertium's MT engine has been released in two open-source packages: lttoolbox (containing all the lexical processing modules and tools) and Apertium itself (containing the rest of the engine); both are available under GNU GPL license. In addition to these programs, open source linguistic data is already available for various language pairs like ca-es, gl-es, pt-es, fr-es, en-es, etc.

With regards to linguistic data for Bologna project, the ES-EN and PT-EN language pairs are at different development stages.

Whereas the ES-EN pair's basic linguistic package has been developed and adaptation work is needed to the linguistic environment, the PT-EN pair is still under development but will be available by the Bologna project start. The adaptation work will be similar to that of the ES-EN pair. It basically involves including proper terminology and adapting some transfer rules in the linguistic data in order to make a translation following a syllabi writing style.

4.3 Cross Language Gateway

Cross Language's translation gateway (Van den Bogaert, 2010) will be extended to provide the collaborative online translation front-end and connect to the MT back-ends. The gateway is a fundamental aspect of the user portal, as it offers a stable, world-ready environment for user interaction with MT connectivity, MT customisation, and experimental MT systems.

Originally developed as a framework for conducting customisation experiments and deploying complex, customised MT routing, the system's scope will be expanded to handle translation requests on the accounting level required by BTS, process the requests and dispatch them to the MT system selected by the BTS user. An elaborate collaboration platform will be added to fit the needs of the Bologna project and the components described in previous sections will be tied together into a single MT application. Currently, the system includes modules for client and user management, accounting and diagnostics. The system is conceived as a web service framework, which allows for easy integration with other components. The key feature of the framework is technology-neutrality in terms of input and output, and in terms of the MT systems used. Incoming documents, in various industry-standard formats, are uniformly handled by various flavours of MT (RBMT, EBMT, and SMT). This allows complex MT set-ups to be configured, modified for improved quality, and used in an easy way.

5 Use Case: Online Study Programme/Syllabus Translation

5.1 Target users and their needs

The primary users of BTS are universities and they will benefit from a fast, low-cost automated translation service enabling them to have English versions of their syllabi and study programmes which is a key requirement in the context of the

Bologna treaty. They will further be in a position to periodically update their existing syllabi and study programmes at a fraction of the time and cost, taking full advantage of leveraging previously produced translations. In effect, BTS provides a solution both for the initial translation volumes and for the maintenance of frequent minor updates. Business users such as companies running educational portals (e.g. mastersportal.eu, findamaster.com, postgraduate.at) will have an opportunity to integrate with the BTS infrastructure resulting in offering multilingual search results to their users. Students will also be able to access the service for translating specific study programmes.

The primary target user group consisting of the universities and the higher education institutes is represented by the Bologna user group, in effect picturing the problem owners and future users.

The Bologna user group will be actively involved in the project from the start in return for a free service throughout the project duration. The members of the Bologna user group will make their previously translated materials available for translation engine training. Moreover they will participate in the user needs analysis and in the evaluation program.

The initial members of the user group cover the languages addressed in this proposal and include: Hogeschool Gent, Belgium; Vrije Universiteit Amsterdam, Holland; Université Charles-de-Gaulle, Lille III, France; University of the Basque Country, Spain; University of Evora, Portugal; Minho University, Portugal; Harbin Institute of Technology, China; Koç University, Turkey; University of Saarland, Germany; University of Eastern Finland, Finland.

5.2 Usage

The user will log on to a portal where he/she can make a request for translating a document via a gateway. Alternatively, upload via API can be done for users who want to integrate BTS into their content repository. For instance, the individual user logs on to the local education database, opens a document and activates the translation service by pushing a translate button in the document interface. Upon translation, a link to a post-edit interface with the translation is sent by email to the user responsible for editing and approving translations. A push on the submit button will send the approved translation to the translation memory as well as to the dedicated location in the database via an API and the gateway in the portal. One or two users may be involved, one activating the translation service and one doing the post-editing.

5.3 Benefits over existing solutions

Available MT services may offer solutions for users looking for an all-purpose gisting translation. The quality levels produced by these services do not, however, propose a solution in the context of course syllabi. Moreover, they do not provide an opportunity to save any corrected or post-edited translations for further re-use. The cost attached to the existing MT solutions prevents them from being considered in this context.

BTS solves the shortcomings of the existing solutions by:

- offering a high quality domain focused MT service
- proposing a self-learning system to continuously reduce the human correction and post-editing effort
- supplying a low cost affordable service
- using technology components which are easily adaptable to additional educational content types
- providing a technology platform catering for rapid language extension
- delivering an open technology infrastructure with easy integration options
- offering various commercial service models including both project-based models as well as integration-based models
- addressing the needs of both institutions and business users

5.4 Potential

According to figures provided by “4 International Colleges & Universities” (www.4icu.org), the number of universities, the prospective users for the proposed translation service, amount to over one thousand in the context of our project covering seven countries. If we were to extrapolate these needs to a European context one may assume a target user group of possibly 2,500 universities. Based on experience with Uppsala University in Sweden, we estimate that major universities with several faculties have between 5,000 and 10,000 syllabi in need of translation (roughly 2,000 for each faculty). For instance, at Uppsala University the total number of unique syllabi to-date amounts to 9,243. As for the medical, technical and agricultural universities,

they are similar to faculties in this respect with some 2,000 syllabi each.

Acknowledgments

This project has received funding from the European Community (ICT-PSP 4th call) under Grant Agreement n° 270915.

References

Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., et al. (2005). An Open Architecture for Transfer-based Machine Translation between Spanish and Basque. *Machine Translation Summit X.* Phuket, Thailand.

Bertoldi, N., & Marcello, F. (2009). Domain Adaptation for Statistical Machine Translation with Monolingual Resources. *Fourth Workshop on Statistical Machine Translation* (pp. 182-189). Athens, Greece: European Association for Computational Linguistics.

Dandapat, S., Forcada, M. L., Groves, D., Penkale, S., Tinsley, J., & Way, A. (2010). OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System. In H. Loftsson (Red.), *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010 (Reykjavíok, 16-18 Aug. 2010)* (Vol. Col. Lecture Notes in Artificial Intelligence, vol. 6233, pp. 121-126). Berlin: Heidelberg: Springer.

Eck, M., Vogel, S., & Waibel, A. (2004). Language Model Adaptation for Statistical Machine Translation based on Information Retrieval. *Proceedings of Language Resources and Evaluation Conference.* Lisbon, Portugal.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). HunPos - an open source trigram tagger . *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 209-212). Prague, Czech Republic: Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005.*
Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL).* Prague, Czech Republic.

Macken, L., Trushkina, J., & Rura, L. (2007). Dutch Parallel Corpus: MT Corpus and translator's aid. *Proceedings of Machine Translation Summit XI.* Copenhagen, Denmark.

Nguyen, B., Hsiao, R., Eck, M., Charoenpornsawat, P., Vogel, S., Schultz, T., et al. (2009). Incremental Adaptation of Speech-to-Speech Translation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Stroudsburg: North American Chapter of the Association for Computational Linguistics.

Nivre, J., Hall, J., & Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, (pp. 2216-2219). Genoa, Italy.

Oflazer, K. (1993). Two-level description of Turkish morphology. *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics* (pp. 472-472). Utrecht, The Netherlands: Association for Computational Linguistics.

Pirinen, T. (2008). Suomen kielen äärellistilainen automaattinen morfologia avoimen lähdekoodin menetelmin. Master's thesis. Helsinki, Finland.

Ramírez-Sánchez, G., Sánchez-Martínez, F., Ortiz-Rojas, S., Pérez-Ortiz, J. A., & Forcada, M. (2006). Opentrad Apertium open-source machine translation system: an opportunity for business and research. *Proceedings of Translating and the Computer 28 Conference.* London.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. *Proceedings of the Conference on Empirical Mehods in Natural Lanugage Processing (EMNLP-96)*, (pp. 133-142). Philadelphia, USA.

Sak, H., Güngör, T., & Saraçlar, M. (2010). Resources for Turkish morphological processing. *Language Resources and Evaluation* , 1-13.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, Italy.

Tiedemann, J., & Nygaard, L. (2004). The OPUS corpus - parallel & free. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).* Lisbon, Portugal.
Van den Bogaert, J. (2010). Technology-Neutral Machine Translation with an Abstracted Technology Stack. *Proceedings of the Workshop on Web Services and Processing Pipelines (WSPP 2010)* (pp. 75-80). Malta: European Language Resources Association (ELRA).

Yuret, D., & Ture, F. (2006). Learning Morphological Disambiguation Rules for Turkish. *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp. 328-334). New York City, USA: Association for Computational Linguistics.