Let'sMT! as a learning platform for SMT

Hanne Fersøe, Dorte Haltrup Hansen, Lene Offersgaard, Sussi Olsen, Claus Povlsen
University of Copenhagen, Centre for Language Technology, Denmark
{hannef, dorteh, leneo, saolsen, cpovlsen}@hum.ku.dk

Abstract

This paper presents an overview of two pilot studies conducted at the University of Copenhagen during the spring of 2013. The studies are based on the use of the Let'sMT! platform, which is also presented in an overview. The purpose of the studies was to investigate whether experiments with the Let'sMT! platform would be adequate for the students as a learning activity during the machine translation components of their courses, and for the instructors to gain experience that would allow them to later integrate the platform properly into the courses. The studies show that the platform is very adequate both for undergraduate and graduate students.

1 Introduction

During the period March 2010 to August 2012 a European consortium developed the Let'sMT! platform with support from the Commission's ICT PSP program[1]. The platform offers user tailored machine translation (MT) and online sharing of training data, and the platform is aimed at professional use in localization.

The main principle of Let'sMT! is to provide a translation platform which allows for user provided content for building MT systems from scratch by using state-of-the art technology in statistical MT. Users may sign up and create their own data repositories and translation systems, or they may use the platform as is with already existing data and systems. A key component that supports the main principle is that it must be very easy, from a technical point of view, to create translation systems and to use them. See a more comprehensive description of the Let'sMT! platform in section 3 below.

After the end of the project, the system is still being maintained. The Centre for Language Technology at the University of Copenhagen (UCPH) wants to use the platform where possible in connection with our core activities as a university, i.e. research and teaching, and to accumulate feed-back from these activities to the consortium partners regarding possible further developments, enhancements, and improvements.

Let'sMT! is a translation platform with a user interface which integrates Moses[2] and saves the user the technical trouble of downloading and installing these tools. From a teaching perspective, this possibility of being able to skip technical difficulties and still be able to teach SMT with a hands-on component is attractive.

We regard our students as advanced MT system users, and we want them in addition to learning about the theoretical aspects of MT, also to try out concrete systems during their course. The purpose of this practical aspect is not to teach them which buttons to press, but for them to learn what machine translation is, which different types of machine translation approaches are available, and how these systems differ from one another in use and output, and what level of knowledge and skills they require from their users. So we want them to study MT in both theory and practice in order for them to become competent users with attractive skill profiles after their graduation, whether they aim at jobs as translators, translation planners, technical support staff, or system developers.

We therefore conducted pilot studies among our students based on the use of the Let'sMT! platform. We wanted to study if the platform is in fact as adequate for learning as we assume, and we also wanted to accumulate some experience for ourselves in order to integrate Let'sMT! as a more comprehensive component in their MT course.

[1] Information and Communication Technologies Policy Support Programme

[2] http://www.statmt.org/moses/, June 18th, 2013

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 353–359.

2 Teaching MT at UCPH

2.1 Existing curricula with MT

The BA course, *Formal Grammar and Machine Translation*, is part of an optional package named *IT and Language* offered to students across all disciplines of humanities. The course starts with an introduction to formal grammars followed by how such a formal description can be used in connection with an automatic transfer-based translation process. The next learning step is a comparison between rule-based and statistically based MT systems, leading again to a more fine-grained walk-through of the principles and concepts behind SMT.

In the walk-through, the basic elements constituting SMT systems are taught, i.e. description of the training data required, the training process, the resulting language models, and finally, the on-line translation process including the decoding algorithm. The course ends with a four hour written exam on a topic given by the lecturer.

The course, *Language Technology 1*, is part of the cross-disciplinary MA study program *IT and Cognition* which is offered to students nationally as well as internationally. The overall MA program combines cognitive and technical courses with a focus on statistical modeling and is designed to provide graduates with competencies adequate for both research and business careers. The main topic of this course is SMT, including how to extend the standard SMT language models by adding relevant linguistic information in order to increase translation quality. Implementation of these additions is an important element of the course.

The main text book for the course is Philipp Koehn's *Statistical Machine Translation* (2010), which is read more or less from beginning to end, i.e. starting with the introduction of basic concepts of SMT and basic linguistic terms, followed by probability theory, various word and phrase modeling types, decoding issues, and evaluation aspects. The exam is the implementation of an addition to the standard model and a report describing the implementation task and the strategy chosen.

2.2 Learning framework

Today's high number of university students with uneven academic backgrounds is challenging traditional teaching and learning approaches at universities.

Recent research focusses on how to maintain high scientific standards and at the same time help the students pass their exams. From a pragmatic angle, the increased competition among universities to attract students has made students' employment prospects a competitive factor. This adds even more emphasis to improving teaching and learning quality.

A relevant theory in this context is outcomes-based teaching and learning (OBTL) (Biggs and Tang, 2009). OBTL is based on three corner stones: Intended learning outcomes, Teaching/learning activities, and Assessment tasks.

According to this theory, a course, ideally, should be based on a description of what the students must learn, i.e. the intended outcomes as documented in the curriculum, how to actually make them learn it, i.e. the lecturer's choice of teaching strategy and learning activities, and how to test their achieved skills, i.e. the assessment tasks or examination.

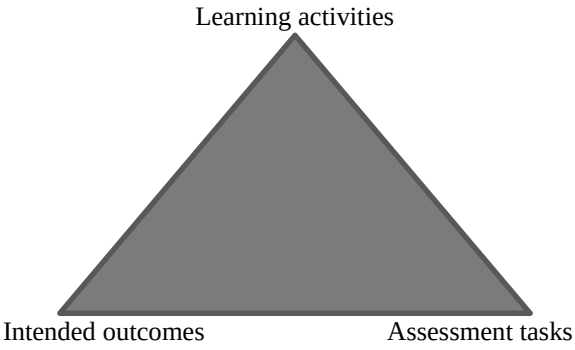

Figure1. The outcomes-based learning triangle

The three corners of the triangle must be aligned, meaning that, for instance, the assessment tasks must be carefully designed to test the intended learning outcomes, as must the teaching and learning activities.

2.3 Learning triangle and pilot study

A very important element in the triangle is the learning activities. In order for the students to acquire the intended learning competencies it is crucial that besides being lectured they also get the opportunity to use their acquired knowledge themselves. By learning the basic principles behind SMT via practical exercises on the Let'sMT! platform, students are being engaged in learning activities that are well aligned with the intended outcome, and particularly the less

academically trained students are expected to benefit from this.

By carrying out the pilot studies before the end of the course, we were able to provide formative assessment, i.e. feed-back during learning. In broad terms, the questions we wanted to get answers to with the pilot study were twofold. In case of the BA students, we wanted to get the students' view of whether they found the platform useful as a cognitive helping device in terms of grasping SMT systems. Concerning the MA students, we wanted to find out whether they would find the platform so easy to use that exercises on the platform would be able to serve as a general introduction to SMT before their course would take them more in depth with technical details. At the time of writing this, the final exam, the summative assessment, has not been concluded, but the feed-back loop works both ways, and we wanted to observe their reaction to the system and discuss it with them in order to learn how best to integrate a Let'sMT! module appropriately into the two courses.

3 Description of the Let'sMT! platform

The Let'sMT! platform allows users to upload their own data to a repository, which converts, stores and handles data in a safe and easy way and prepares the data for training of standard SMT engines (Vasiljevs et al 2012, Tiedemann et al. 2012). The uploaded data and the trained SMT systems cannot be downloaded from the platform again, they can only be used within it. This is a security measure implemented in order to protect uploaded data, particularly private user data against unauthorized use. For testing purposes and for minor translation tasks, an online translation web interface is available, while the platform allows for integration with the translation memory systems SDL Trados and MemoQ via a plug-in for professional use scenarios.

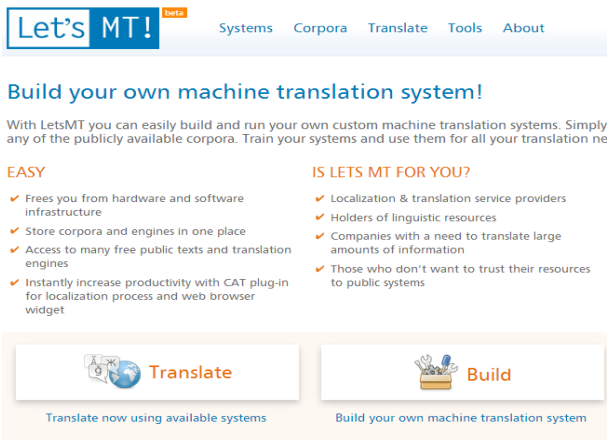

Figure 2. The Let'sMT! Platform (https://www.letsmt.eu)

3.1 System training

From an easy-to-use web interface, registered users can configure SMT engines based on a combination of resources uploaded for public use on the platform or resources uploaded by the user him or herself for private use only. The data can be parallel as well as monolingual and from different subject domains.

Cloud-based system training is then carried out based on the Moses SMT software (Koehn et al. 2007) using in-domain and out-of domain data weighting as described in (Koehn and Schroeder, 2007). Finally, the trained systems are evaluated on either a user defined evaluation corpus or on a system generated evaluation corpus using both BLEU, NIST, TER and METEOR evaluation metrics.

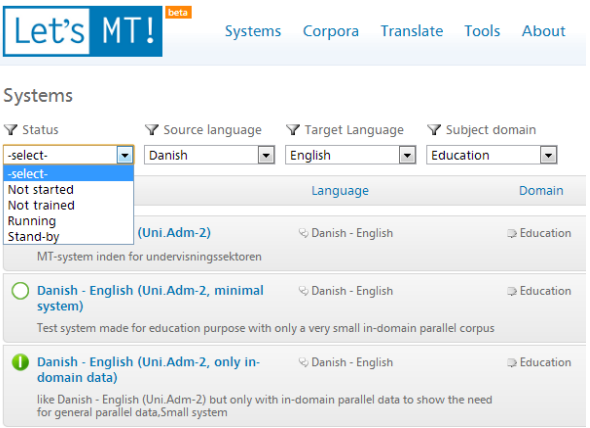

Figure 3. Overview of trained systems (https://www.letsmt.eu/ Systems.aspx)

3.2 Data

The Let'sMT! platform lets users train domain-specific systems based on data uploaded to the Let'sMT! resource repository. The data available in the repository consists of the large and well-known publicly available corpora e.g. Europarl, DGT-TM Acquis Communitare and the Opus corpora, all of which are often used for SMT systems. In addition to these resources, domain-specific data for several under-resourced languages are available. Currently the data covers 9 different subject domains: *Biotechnology and health*, *Education*, *Electronics*, *Environment*, *Finance*, *IT*, *Law*, *Tourism*, and *National and international organizations and affairs* with focus on the 10 languages within the Let'sMT! consortium: *Croatian*, *Czech*, *Danish*, *Dutch*, *Estonian*, *Latvian*, *Lithuanian*, *Polish*, *Slovak*, *and Swedish.*

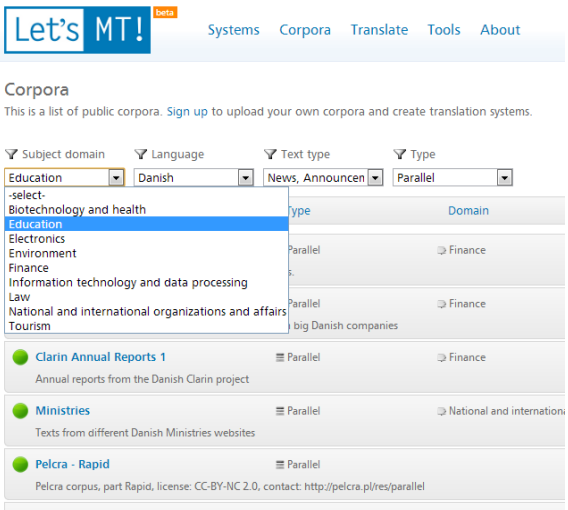

Figure 4. Selected corpora (https://www.letsmt.eu/ Corpora.aspx)

3.3 Student use

In this section we will focus on the features that make Let'sMT! particularly well suited for active student participation.

Usually SMT hands-on excercises in teaching requires scripting and programming skills, which are not the main focus for e.g. BA students from language studies. Let'sMT! lets the students get a hands-on experience, both with the training process and with the use of aligned data without time consuming technical obstacles.

It is important to teach the students the role of data, because the linguistic knowledge in an SMT system is built on the training data. Often the large, well known corpora such as Europarl are used as training material, but the use of these corpora does not give a student a good understanding of the balance between data quality and data quantity. The selection of proper training resources is especially important for under-resourced languages and for translation of texts from specific domains. The variety of corpora at the Let'sMT! platform facilitates this selection.

Let'sMT! offers free on-line user registration, which makes it possible for the students to sign up during the course. As users of the platform they can translate texts with the available public systems, they can train systems using less than 2M parallel sentence as training material, and they can also translate using their own systems. However, a free user account does not allow for upload of corpora; this activity has to be executed from a license-fee account. The licensed teacher of the SMT course therefore has to upload extra corpus data for the students when relevant.

Upload of parallel corpora as training material is easy for a large number of formats. The platform offers automatic format conversion and alignment of the data, and the user can inspect potential warning messages from the upload process and see the sizes of the resulting aligned corpora.

In the "Create System" web interface of the platform, the workflow for training an SMT system is broken down into a number of steps. This modular way of specifying the needed information for the training tasks leads the students through the process in a structured manner. Help texts are available and the user gets hints and feedback if chosen options or corpora diverge from recommended usage.

The progress of an ongoing training process can be followed by inspecting a training chart. This chart gives information about both the complexity of the training process and about the flow of processes. In addition, it shows the current state of the training process.

To evaluate the quality of the trained systems, one can either upload an evaluation corpus, use an existing evaluation corpus on the platform, or let the platform extract an evaluation corpus from the training data. As the platform automatically calculates scores for the most common evaluation metrics, students can investigate the metrics and the translation quality by downloading the evaluation corpus and the resulting translation.

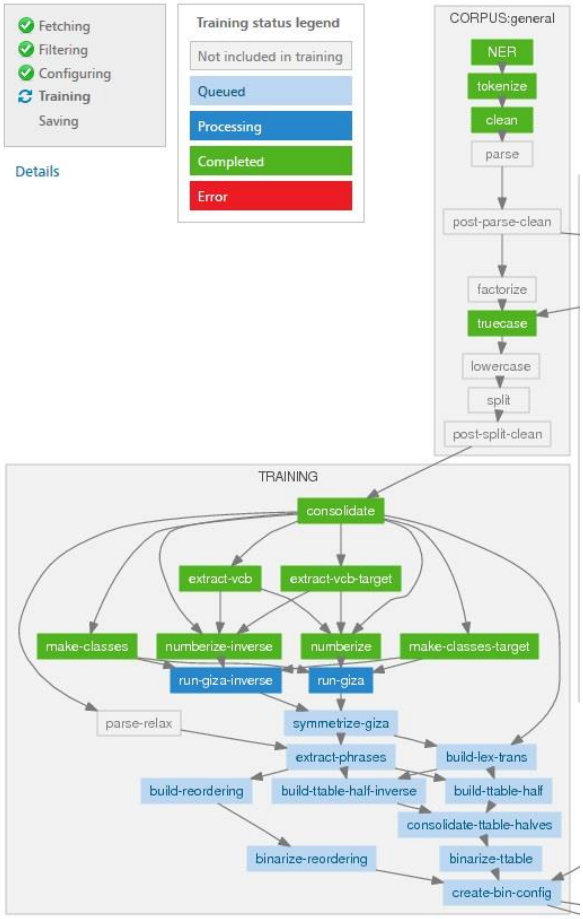

Figure 5. Part of training chart

Finally, the options to integrate Let'sMT! into TM systems can be used to show the students an example of a professional translation workflow.

4 The pilot studies

The pilot studies take as their points of departure the current BA and MA curricula as described above, the student body in the courses of the spring semester of 2013, and the practical possibility of accommodating such a study into the on-going teaching plan.

4.1 BA students in class with Let'sMT!

An important element in the BA course was the goal of teaching the students the type of data that is required to develop an SMT system. To meet this goal Lets'MT! excercises would be a useful learning activity for the students to engage in and acquire new knowledge. By using Let'sMT!, they were able to go through the training process of SMT by following the stepwise guide in the "Create system" web interface, cf.3.3 above. The training data of the small SMT system trained consisted of about 20.000 bilingual sentences (English and Danish) and 20.000 monolingual sentences all within the subject domain, *Education*.

Later in the course, the output from this SMT system served as part of a comparative analysis the aim of which was to get the students to understand the interdependency between the types of training data, the types of input text, and the translation quality. Here again the platform formed the basis for a practical learning activity. As the initial part of the exercise, an input text from the subject domain of *Education* was translated using the above mentioned in-domain SMT system on the Let'sMT! platform. Then the students took the same input text and translated it by using Google Translate, which is characterized by being general, i.e. subject domain neutral.

Finally the two translation results were compared. For illustratives purposes see the following translation of a sample sentence from the input text:

The Danish input sentence: DA: Studieordningen for Bachelortilvalg i vikingestudier (2011) er udarbejdet af Studienævnet ved Nordisk Institut
Translation results made by the SMT system via Let'sMT! (in domain): EN: The academic regulations The academic regulations for the Bachelor's Supplementary Subject in Viking Studies (2011) were prepared by the Board of Studies at the Department of Nordic Studies
Translation results from Google Translate (out-of-domain)[3]: EN: The curriculum for the Bachelor Options in Viking Studies (2011) prepared by the Board of Studies at the Nordic Institute
The correct translation: EN: The academic regulations for the Bachelor 's Supplementary Subject in Viking Studies (2011) were prepared by the Board of Studies at the Department of Nordic Studies

Table 1: Example translation from the BA class.

Not surprisingly the in-domain system has a better performance with respect to finding the correct domain specific terms, although, due to lack of coverage, it sometimes fails in terms of unexpected translations. As the example illustrates, the somewhat small but in-domain SMT

[3] http://translate.google.dk/ (accessed, May 15, 2013)

system performs better in terms of translation quality than translations generated by domain-neutral system such as Google Translate. Especially concerning the translations of technical terms, the in-domain system proved superior.

This experiment, easily carried out by using Let'sMT!, made it transparent and comprehensible to the students that the type of training data and the text type to be translated are interrelated when it comes to translation quality.

4.2 MA students in class with Let'sMT!

The MT course of the advanced MA students was planned long before the idea of this study was conceived. For this reason it was not possible to integrate learning activities based on the Let'sMT! platform into the study plan in the same way as was the case for the BA students. Instead, we planned to organize a focus group meeting where Let'sMT! would be introduced, tried out, and discussed. The focus group meeting took place at the end of the course where the students already had acquired in-depth theoretical and practical learning goals in SMT.

The meeting started with a presentation of the Let'sMT! platform where the MA students were guided through all the different aspects of data upload (including alignment), the steps of system training (illustrated by a dynamic training chart, see figure 5), and the automatic evaluation of the results.

After this introduction, the students were instructed to make hands-on experiments with the platform on their own, with the possibility of asking questions to the instructor in the process. Finally a discussion took place between the students and the teachers about the platform and its potential use and usefulness in future MA courses.

Since the MA students were already familiar with SMT theory and hands-on processes, the outcome of the focus group meeting was of another kind than what was the case with the BA study.

5 Outcome of the pilot studies

In order to be able to evaluate what came out of the pilot studies, i.e. the extent to which the use of the Let'sMT! platform had contributed to a better learning process for the students and thus to their better understanding the principles behind SMT, it was decided to conduct a survey in each of the student groups in order to collect their feed-back.

5.1 Outcome of the BA pilot study

The survey served as part of the formative assessment. The students were asked to give their assessment of how useful the Let'sMT! platform had been as a cognitive helping device in terms of grasping the principle of SMT systems. The questions were about understanding the types of training data required, the correlation between training data and input text, and finally how user friendly the Let'sMT! user interface appeared to be. The assessment was given in the form of scores on a scale from one to five.

The average score was close to four for all the questions asked. Based on this feedback, we can conclude that using the Let'sMT! platform as an instrument to involve the students in a learning activity, appears to be a great advantage.

The students were also encouraged to give comments on how the platform could be improved. Not surprisingly, several of the comments had to do with how to make the user interface look more like Google Translate.

5.2 Outcome of the MA pilot study

Like the BA students, the MA students were also introduced to a survey in order to express their assessment about the potential usefulness of Let'sMT! in future MA courses.

We wanted to know whether Let'sMT! could be of any help to advanced students in facilitating their understanding of SMT and the different steps of system training. They were therefore asked whether acquaintance with the Let'sMT! platform would have been helpful during the course for a better comprehension of SMT.

We also wanted to find out whether the advanced students would find Let'sMT! appropriate for maing their own experiments, e.g. training of systems with different data sets and languages. Finally, we wanted to see whether the MA students would find Let'sMT! suitable for evaluation tasks with different evaluation metrics.

In the survey they were also asked to report on the shortcomings and limitations of the platform from a pedagogical and educational point of view and to give suggestions for future use of the Let'sMT! platform in the different parts of their study program.

The students reported that Let'sMT! worked well as an initial introduction to SMT and would have worked well at the beginning of their course. They saw it as an easy way to follow the different steps of how an MT-system is created and works, and they found that especially the

training chart (see figure 5) gives a good and pedagogical overview. They also found it easy to train a system with the guidance offered on the platform.

The students found it very useful to be able to try out the training and use of systems with different types of data and language pairs for standard experiments. They saw this as a fast way of trying out different corpora to see which ones to use in a project.

However, they complained about the system being a 'black box' meaning that it was hidden for them what was going on inside. Advanced students need more technical information and they need to be able to go into details. E.g. one of the experiments that the students have to do during their course is factored translation, which they cannot do in Let'sMT!. The students also found the system a bit slow for classroom use.

In addition to these observations, the students saw the great usefulness in the easy access to a range of evaluation scores. Some argued that they would definitely use this function in future work.

They also had suggestions for platform extensions that would support the use of the platform in a learning context.

They suggested that the platform should be extended with more technical documentation of the training chart explaining the various steps and linking more directly to publications describing the implemented architecture. They also suggested to extend the guidance to what is necessary when training a good MT system, for instance information about the importance of in-domain corpora. Other suggestions concern having access to a list of the most used systems, and the possibility of creating a group for the license-free accounts in an educational context, where access to each others' systems would be very useful.

6 Conclusion and future prospects

Summing up for the BA students, they particularly benefitted from the platform as a practical, easy-to-use cognitive helping device, and an obvious future prospect is to even further integrate the Let'sMT! platform into the curriculum and course plans.

Summing up for the MA students, they particularly saw their use of the platform as a potential learning activity in the initial introductory phase of their course, where the different steps of SMT system creation and the training chart offered shortcuts to the more in-depth technical details that they would go into later in their course. So for these students the indication is also clear; integration into their study plan will support their SMT learning.

We can conclude from the pilot studies and from the surveys that although Let'sMT! was developed for professional localization scenarios without any views to teaching contexts, it is in fact very adequate as a learning platform.

References

Biggs, John, and Catherine Tang, 2009. *Teaching for Quality Learning at University*. Society for Research into Higher Education & Open University Press.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine trans¬lation. In *Proceedings of the Second Workshop on Statistical Machine Translation,* pp. 224–227, Prague, Czech Republic.

Offersgaard, Lene, and Dorte H. Hansen, 2012, SMT systems for less-resourced languages based on domain-specific data. In *Proceedings of The 5th Workshop on Building and Using Comparable Corpora*. European language resources distribution agency, pp. 75-80.

Tiedemann, Jörg, Dorte H. Hansen, Lene Offersgaard, Sussi Olsen, and Matthias Zumpe. 2012. A Distributed Resource Repository for Cloud-Based Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European language resources distribution agency, Istanbul, Tyrkiet, pp. 2207-2213.

Vasiļjevs, Andrejs, Raivis Skadiņš, and Jörg Tiedemann. 2012. LetsMT!: a cloud-based platform for do-it-yourself machine translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (ACL2012), System Demonstrations, pp. 43-48. Jeju, Republic of Korea.