1. PUBLISHABLE SUMMARY

Summary of the context and overall objectives of the project (For the final period, include the conclusions of the action)

Rapid translation between European languages is a cornerstone of good governance in the EU, and of great academic and commercial interest. Data-driven approaches to machine translation based on machine learning techniques are widely used and constitute the state-of-the-art. The basic knowledge source is a parallel corpus, texts and their translations. For domains where large parallel corpora are available, such as the proceedings of the European Parliament, a high level of translation quality is reached. However, in countless other domains where large parallel corpora are not available, such as medical literature or legal decisions, translation quality is unacceptably poor. Given the strong demand for automatic translation capabilities this is a problem of critical importance.

We are working on solving two basic problems of knowledge acquisition for machine translation. The first problem is determining how to benefit from large out-of-domain parallel corpora in domain-specific translation systems. The second problem is mining and appropriately weighting knowledge available from in-domain texts which are not parallel.

Our work will lead to a break-through in translation quality for the vast number of domains with less parallel text available, and have a direct impact on companies providing translation services. The academic impact of our work will also be large because solutions to the challenge of domain adaptation apply to all natural language processing systems and in numerous other areas of artificial intelligence research based on machine learning approaches.

Work performed from the beginning of the project to the end of the period covered by the report and main results achieved so far (For the final period please include an overview of the results and their exploitation and dissemination)

In the first year of the project, we carried out work on improving translation to morphologically rich languages using classifiers, which is critically important to domain adaptation as it allows the integration of lemmas (given by a user or automatically mined from the web) with proper inflection into the translation process. This work was integrated into the Moses open source statistical machine translation system which is widely used in both academic and commercial environments. We had important followup papers on this work at the beginning of the second year, in addition to carrying out work on better linguistic modeling and on modeling transliteration (the

process of mapping proper nouns like a person's name from one script to another, e.g., as when translating a name from Arabic script to English). We also studied neural machine translation.

In the second year of the project, we completely switched to neural machine translation, a new technology overcoming some limitations in the previous state-of-the-art (which was phrase-based statistical machine translation). We carried out important work here on both inflectional generalization and improving linguistic representation, as well as on fast training algorithms. We participated in an important machine translation community shared task, and had excellent results (with a particular highlight being having the best English to German system for news translation according to human judgments of the translation output).

In the third year of the project (which we are now in the middle of), we developed new technology for automatically finding the translation of terms which are not in our parallel training data. We also showed how to leverage large in-domain corpora in tasks like this one and other tasks in natural language processing. An interesting result of this work is that we have very good performance on detecting the sentiment of Spanish tweets without using Spanish language training data (i.e., only using English language training data).

We have also begun work on two important new areas of research, which we had not previously planned to address. The first is on training machine translation systems without the use of any parallel data. This is an exciting development that will allow us to address translation tasks which we were not able to previously consider due to lack of training data. The second is that we have carried out research on document translation, which allows us to use information from the full document in the translation process, allowing us to better model, e.g., an ambiguous word in terms of the full context of the document rather than only using the sentence the word occurs in.

Progress beyond the state of the art and expected results until the end of the project

We have improved the state-of-the-art in phrase-based statistical machine translation by incorporating classifiers for dealing with rich morphology, and by studying linguistic and script-related problems.

We have also improved the state-of-the-art with respect to using linguistic information in neural machine translation, using bilingual word embeddings for finding the translations of unknown terms from comparable corpora (such as Wikipedia), and cross-language adaptation of classifiers, particularly in the unsupervised case.

In the next 2.5 years, we will scale our approaches for mining terms

and parallel sentences to web-scale. We will show how to use mined terms and parallel sentences in neural machine translation systems to solve the out-of-vocabulary problem which occurs when carrying out domain adaptation. We will complete the creation of a document translation system, which uses document-level context in translation (rather than considering each sentence in isolation), and show how it models domain. Finally, we will create state-of-the-art unsupervised machine translation systems, i.e., systems which do not require any parallel training data, which we will apply to domain adaptation tasks where we have no in-domain data.

Address (URL) of the project's public website

http://www.cis.uni-muenchen.de/~fraser/dasmt.html