ACCEPT

COOPERATION

Work package 4: Improving SMT

Description of work

Description of work

Task 4.1: Baseline machine translation systems (UEDIN, SYMANTEC, month 1-3)
Given the Moses toolkit and initially available data resources, we will build baseline machine translation systems for all domain settings of interest at the beginning of the project.

Task 4.2: Domain adaptation methods (UEDIN, SYMANTEC, month 4-24)
In the use cases tackled by the ACCEPT project specialised domains are targeted, while typically large amounts of out-of-domain data are available. There has been some work on balancing in-domain and out-of-domain data, such as by weighting feature functions associated with different translation models, use of interpolated or multiple language models, etc., but given the scope of the problem, it has not been sufficiently addressed. We will work on new interpolation techniques inspired by interpolation in language models, discriminative training methods, and other approaches to develop novel domain adaptation methods. We will explore a number of approaches: (a) linear combination of in-domain (InD) and out-of-domain (OutD) models, (b) backoff methods to use OutD models only for rare and words and phrases (in InD), (c) use of features for each phrase pair indicating relative frequency in InD and OutD in discriminative training, (d) sentence-level scoring of training data based on similarity to test data, (e) subsampling of OutD data with preference to sentence pairs with source similarity to test set, (f) adaptation methods applied to InD monolingual source data.

Task 4.3: Linguistic back-off (UEDIN, UNIGE, acrolinx, month 4-24)
In the EuroMatrix project [Koehn and Hoang, EMNLP 2007], we developed the so-called factored translation model framework, which allows for various decompositions of word and phrase translation. For example lemmas and morphological properties can be translated separately, as opposed to just translating the surface form as in the standard models. In the ACCEPT project, in cases of morphologically rich language pairs with little in-domain parallel data (or any parallel data), we will build on this framework to provide robust back-off from the usual translation of surface forms to the translation of word stems. The model will allow us to translate the surface forms of frequent words and phrases, but to use a decomposed or synonym translation for rare or unknown words. The correct output word forms will be generated by a generation module.

Task 4.4: Exploitation of usage data (UEDIN, month 13-36)
We are developing machine translation systems for a specific purpose, i.e., to leverage the machine translation output to produce high-quality translations with the aid of human pre- and post-editors. The edits by human editors to the machine translation output are a valuable data source to determine where the system failed. We will develop methods to exploit this data during the system optimisation (tuning) stage to target system performance to overcome its observed faults.

Deliverables	Delivery date	Description
D 4.1	PM 3	Baseline machine translation systems.
D 4.2	PM 24	Report on robust machine translation: domain adaptation and linguistic back-off.
D 4.3	PM 36	Report on improved machine translation by exploiting post-editing data.