Integrated Language Technology as part of Next Generation Localisation

Julie Carson-Berndsen[1], Harold Somers[2], Carl Vogel[3], Andy Way[2]
Centre for Next Generation Localisation
[1] School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
[2] School of Computer Science,
Dublin City University, Glasnevin, Dublin 9, Ireland
[3] School of Computer Science and Statistics,
Trinity College, College Green, Dublin 2, Ireland
www.cngl.ie
hsomers@computing.dcu.ie; away@computing.dcu.ie;
vogel@cs.tcd.ie; julie.berndsen@ucd.ie

Abstract

This paper describes one component of a large research project involving industry-academia collaboration between four Irish universities and nine Irish and multinational industry partners, all collaborating to develop 'Next Generation Localisation'. The project as a whole is described by van Genabith (2009); the current paper focuses on on the role in the project of state-of-the-art language technology including text and speech processing, and machine translation (MT) in its various forms. In this paper, we describe the basic and innovative research approach to integrating language technology into the overall design. We will describe research in the areas of MT, speech technology and text analytics (TA), and ways in which these three are closely integrated with each other.

Keywords: *language technology, machine translation, localisation, standards, evaluation, speech technologies, crowd sourcing, translation memories, post-editing*

1. Introduction

As mentioned by van Genabith (op. cit.), Next Generation Localisation seeks to address current problems of increased volume, access and personalisation. Regarding volume, increased automation is the only viable approach to meet the challenges posed by the spiralling amount of material to be localised worldwide. Automation is particularly relevant to the core task of localisation, namely translation. Since its inception, the localisation industry has been highly computerised, linking up and supporting teams of human translators, localisation project managers and customers with Translation Memory (TM) and terminology management systems, electronic dictionaries, translators' workbenches and localisation workflow management and quality assurance systems. However to date, core language technologies, in particular MT, have surprisingly been incorporated only sparingly into the localisation process. This is largely due to early unrealistic expectations (on the part of the users/Localisation Industry), unwarranted promises (on the part of MT researchers and developers) and the ensuing disappointment and reluctance to invest in MT technology in the localisation workflow. Right now, this situation is changing dramatically: data-driven and machine-learning (ML)-based language and MT technologies are revolutionising MT research, achieving analysis and translation quality, coverage and robustness at a cost previously unimaginable. The proper place of MT is being investigated systematically: MT must be integrated into a translation and post-editing workflow together with human translators to tackle volume and to save costs; task-dependent configurable systems covering the complete spectrum of fully automatic raw translation, human-aided machine or machine-aided human translation with their associated different levels of translation quality can each play a role in the localisation process. Novel ML-based language technologies can automatically provide metadata annotations (labels) to localisation input to automate localisation standardisation and management. And progress in MT evaluation can measure localisation output and automatically cost localisation input. All of these factors determine the approach to MT research in this

project, with the aim of advancing basic research in integrated MT-centred language technologies with the following aims:

(a) to improve MT engines to achieve increased quality and hence automation of translation in the Next Generation Localisation 'Factory', in which technologies are integrated into workflows and complex information-technology-based software systems (van Genabith 2009)
(b) to develop technology which will automatically annotate localisation input with standardised localisation metadata so as to automate localisation workflows,
(c) to develop novel MT evaluation methodologies, and
(d) to evaluate the impact of automation in the localisation workflow.

Regarding **access** and **personalisation**, we see integration of MT with speech technologies as a crucial step. Small screen and non-keyboard-based devices (mobile phones, PDAs) increasingly support affordable, pervasive, on-the-move access to globally available multilingual digital content. Novel speech interfaces will be essential to compensate for the main limitations of such devices, to provide support for 'eyes-busy, hands-busy' scenarios as well as for handicapped users (e.g. the blind). Unfortunately, to date, standard localisation has made little provision for the optimal adaptation of content to such devices, even though speech recognition and synthesis can potentially extract and provide information highly relevant to personalised delivery of and access to digital content. The present project thus supports both text- and speech-based mobile access to and delivery of multilingual information as well as personalisation of information. In order to achieve this, fundamental problems of speaker and language dependence of current state-of-the-art speech recognition systems (amplified in the multilingual, mobile, instant and online Next Generation Localisation scenario) need to be solved and a proper integration of speech and translation technologies needs to be provided: while there are striking similarities between state-of-the-art ML-based approaches to speech and MT (in terms of the underlying technologies both for statistical and example-based methods), a fully integrated speech and MT system which can share and exploit information provided by both components in a mutually beneficial way is still lacking. Basic research in speech-based interfaces has the following aims:

(a) to produce tightly coupled speech and MT technology capable of mutually and maximally exploiting information provided by each component,
(b) to reduce speaker and language dependence of speech technology through the use of linguistically motivated, ML-based hierarchies,
(c) to extract information (such as gender, age, emotion) automatically from speech input relevant to personalisation and generate personalised speech output, and
(d) to evaluate the impact of speech interfaces in the context of localisation.

All of these goals are supported by research on Text Analysis (TA), for which two core tasks are defined, namely automatic annotation of localisation data with metadata, and text classification. Reliable automatic multilingual text classification is required to tune suites of novel MT and speech processing systems to text-type and genre. Automatic labelling is required to annotate multilingual input with standardised metadata to automate localisation workflows and to annotate multilingual corpora with dependency information to induce novel probabilistic transfer-based MT systems and to provide syntactic information for syntax-boosted MT systems.

In the remainder of this paper, we will provide further details of the research work already started, and planned for the future, to address these goals. The work represents a collaboration between five groups. At DCU, the research is carried out within the Language and Intelligence Research Group in the School of Computer Science and in the Centre for Translation and Textual Studies in the School of Applied Language and Intercultural Studies. Within Trinity College, research is conducted through the Phonetics Laboratory in the Centre for Language and Communication Studies, and by the Computational Linguistics Group, part of Intelligent Systems in the School of Computer Science and Statistics. At UCD, research is undertaken by the MUSTER group in the School of Computer Science and Informatics. Languages already addressed by the MT and MUSTER groups include French, German, Chinese, Arabic, Japanese, Polish and Spanish, usually paired with English, and new personnel mean that we may address in addition Turkish, Hindi, Bengali, Irish and Irish Sign Language. Of course, not all languages or langauge pairs/directions are covered to the same degree. Work on stylistic analysis at Trinity College includes English, Danish, Finnish, Norwegian and Russian.

2. Machine Translation

A huge demand for MT exists already: web service providers process millions of requests for automatic translation every day. Until recently, the service offered by Google[1] was powered by a version of the successful Systran system, which was also behind the well-known Babelfish service[2], among others. However, this older rule-based and hand-crafted technology is in the process of being replaced by a new generation of data-driven and ML-based statistical (SMT) systems, and these technologies have dominated MT research for at least the last ten years. Nevertheless, it is being recognised that purely statistics-based MT systems will not deliver the high level of quality demanded by some applications, including some localisation scenarios. Accordingly, current research seeks to integrate better linguistic processing into statistical approaches. This is not generally felt to indicate a return to rule-based approaches, but to hybrid designs where the appropriate linguistic knowledge is extracted from corpora by statistical and other means, and integrated into MT systems which are still nevertheless driven by the basic approach that was first introduced in the 1990s, and is just now reaching a maturity based on the huge amounts of research effort dedicated to it. Our belief is that MT translation quality will be further improved through fundamental advances resulting from combining the Example-based (EBMT: Nagao 1984, Somers 1999, Carl & Way 2003) and Phrase-based SMT (PB-SMT: Marcu & Wong 2002, Koehn et al. 2003) paradigms, from the introduction of syntactic information in EBMT (e.g. Hearne & Way 2006) and SMT (Chiang 2005) to better capture global reordering, and from fine-tuning ML-based systems to text-type and genre (e.g. Ueffing et al. 2007).

A large percentage of the research in this area is dedicated to the development of core MT engines. Building on previous work by the research team, we focus on six key challenges for MT research.

2.1.Syntax-based SMT

In contrast to work which has shown that SMT performance degrades when seeded with more syntactic units (cf. Koehn et al. 2003), we have shown in previous work that incorporating models of syntax into SMT systems can improve translation quality. There are two aspects to this work:

(a) incorporating syntax in the source language;
(b) incorporating syntax into the target-language model and the translation model.

With respect to the first of these, in Stroppa et al. (2007), we demonstrated that the performance of the state-of-the-art PB-SMT system Moses (Koehn et al. 2007) can be improved significantly when context-informed features of the source language (here, neighbouring words and their part-of-speech categories) are used. These context-informed features are integrated directly into the original log-linear framework (Och & Ney 2002), while benefiting from the existing training and optimisation procedures in standard PB-SMT.

Building on this previous work, in Haque at al. (2009a) we showed that using 'supertags' (Bangalore & Joshi 1999, Clark & Curran 2004; see also next paragraph) can provide still further gains, while in Haque et al. (2009b), we demonstrated that source-language dependency information can also improve target-language output. This work is ongoing with international collaborators from the University of Tilburg, using their suite of memory-based classifiers (Daelemans & van den Bosch 2005).

With respect to the second research thrust in this track, in Hassan et al. (2006) we incorporated into Och & Ney's (2002) log-linear PB-SMT framework a novel language model based on supertags as well as a translation model whose target side included supertags to improve the BLEU score (Papineni et al. 2002) on a range of tasks, and for different language pairs. The intention behind this research thrust is to build on our previous work to incorporate target-language and bilingual constraints into the range of MT systems being developed in CNGL at DCU, in conjunction with our industrial partner IBM.

2.2. Hybrid MT systems

A second challenge is to merge SMT and EBMT into improved hybrid systems: recent ground-breaking work by Way & Gough (2005a) and Groves & Way (2005) extended previous approaches to hybrid MT by showing that combining EBMT and SMT subsentential alignments in novel 'Example-based SMT' and 'Statistical EBMT' systems improved translation quality over baseline EBMT and SMT systems. We are now porting these insights to new domains and language pairs, and integrating a novel EBMT decoder (Groves, 2007). Further novel research involves combining sets of automatically

1 http://translate.google.com
2 http://babelfish.yahoo.com

induced generalised templates clustered around content and closed-class words in EBMT (Brown 1999, Way & Gough 2003), where such templates are commonly used to increase coverage and translation quality, and in SMT, where such templates are not used at all. This work is being undertaken in close collaboration with our industrial partner Traslán.

A further research thrust in this area extends our previous work (Way & Gough 2005b) to build large-scale Controlled Translation systems (O'Brien 2003, O'Brien & Roturier 2007). This work is being undertaken in close collaboration with our industrial partner Symantec.

2.3. Scaling up

The third challenge is automatically scaling more linguistically sophisticated systems to larger amounts of training data. Our Data-Oriented MT (DOT) systems (Hearne & Way 2006) have already been shown, with limited amounts of appropriately annotated training data, to outperform state-of-the-art SMT systems. We are now scaling up by at least two orders of magnitude the amount of training data used by these systems, and developing novel scoring methods for DOT (Galron et al. 2009) to enable closer comparison with mainstream PB-SMT systems, especially in the parameter estimation stage (cf. Och 2003).

At the same time, we are automatically inducing the complete set of resources for large-scale probabilistic transfer-based MT from parallel texts with our treebank-based multilingual, probabilistic LFG (Lexical-Functional Grammar) parsers/generators (Cahill et al. 2004, Cahill and van Genabith 2006, Graham et al. 2009).

2.4. Tuning to text-type and genre

The fourth challenge is to tune ML-based MT to text-type and genre. Optimal lexical selection and syntactic choices can only be achieved if an MT system is tuned to a particular domain. To date, there has been surprisingly little research on tuning ML-based MT technology to particular domains, text-types or genre. Given the quality and range of training material provided by industrial partners in this research, including computer systems, security, office applications, primary and secondary legislation, and printing, we are able to tune a suite of MT systems (SMT, EBMT, hybrid and probabilistic transfer-based) to domain, text-type and genre (cf. Haque et al. 2009c), according to a classification model using classifiers from the TA research reported below, and to investigate how much supplemental genre-typical training material is required, and to investigate training on comparable, rather than full parallel text, resources. This work is being conducted with many of our industrial partners, including Symantec, Traslán, DNP, and VistaTEC.

2.5. Alignment models

Our fifth challenge is to develop novel alignment models for MT technologies based on a range of types of training data: trees, strings, dependency structures etc. This work has two main foci: word alignment and phrase alignment.

Our previous work (Ma et al. 2007, Ma & Way 2009) has shown that new models of word alignment can be developed which outperform state-of-the-art methods (Och & Ney 2003, Deng & Byrne 2005, 2006). Based on this previous work, in Lambert et al. (2009), we demonstrated that tuning word alignment on an extrinsic task (MT, here) rather than intrinsically (compared to a 'gold standard' set of word alignments) can improve translation quality on a range of language pairs and on different domains. With respect to the induction of phrase pairs, our prior work has shown that statistical models of translation can be improved by incorporating example-based source-target chunks (Groves & Way 2005, 2006), as well as pairs derived using dependency (Tinsley et al. 2008) and constituency trees (Tinsley & Way 2009). This work has recently been scaled up by two orders of magnitude in Srivastava & Way (2009), as well as extended to incorporate phrase pairs induced from head-percolation information (Magerman 1995).

Ultimately, for both word and phrasal alignments, we would like to develop a novel general alignment model which, in abstracting away from the surface differences in annotation, is capable of inducing subsentential alignments over source-target pairs no matter which type of annotation is provided.

2.6. Evaluation

The final challenge in this part of the research programme is to develop improved automatic MT evaluation technology. This is a key component of MT research and development. To date, most automatic evaluation methods are based on string matching (e.g. BLEU, Papineni et al. 2002), and have been shown to penalise legitimate lexical and syntactic variation (Callison-Burch et al. 2006). In order to account for such variation, such methods require multiple references, which are time-

consuming and expensive to construct. Based on our previous work (Owczarzak et al. 2006), we are developing a novel dependency-based MT evaluation technology that automatically accounts for both lexical and syntactic variation. In He & Way (2009a), we showed that the labelled dependencies in Owczarzak et al. (2006) could be learned by ML techniques with a corresponding increase in correlation with human judgements. In He & Way (2009b) meanwhile, we demonstrated that parameters trained on one metric (BLEU, say) using the method of Och (2003) may not lead to optimal scores on the same metric, especially where only a single reference translation is provided. Furthermore, combining different evaluation metrics not only reduces any bias related to any one particular metric, but also gives better translation quality than tuning on any standalone metric.

Putting all these together, we aim to produce a suite of novel core MT engines for the purposes of localisation as widely understood by the project as a whole, including PB- SMT systems with integrated syntactic models, improved models of EBMT, novel hybrid data-driven EBMT-PB-SMT systems, novel discriminative and controlled EBMT engines, large-scale DOT systems, novel transfer-based MT engines induced from automatically labelled parallel corpora, systems tuned to text-type and genre, novel aligners, EBMT decoders, and automatic MT evaluation methods.

3. Speech technologies

Flexible, non-keyboard-dependent, on-the-move voice access and response is a core enabling technology for intelligent access to digital content. Speech interfaces to mobile devices are essential in 'eyes-busy, hands-busy' scenarios. In the multilingual application scenario addressed by the CNGL research project, a tight integration of speech technologies and MT is imperative to achieve optimal results. In order to address the project goals of volume, speech recognition and synthesis systems need to deal with potentially an unlimited vocabulary, with multiple (and non-native) speakers and with multiple languages. Speech carries information on multiple levels. For example, personal information such as gender and age is communicated by voice characteristics; prosody and voice quality carry crucial grammatical information; emotional state or mood is communicated by prosody and tone of voice; sound qualities and systematic patterns distinguish between native and non-native users. Such factors are required to support personalisation.

Within the context of the Next Generation Localisation project, the key research challenges for speech technology are the design, development and evaluation of a new breed of robust and scalable automatic speech recognition (ASR) and speech synthesis engines which overcome problems associated with unrestricted domains and speaker types and which facilitate porting of the technologies to new languages (Carson-Berndsen & Walsh 2005). We are developing intelligent engines which utilise the multiple levels of human expressive speech, which are being integrated in a novel way with the MT models described above to facilitate speech-to-speech MT, text-to-speech MT and speech-to-text MT. By defining the experimentation domain for tight coupling in terms of an annotation hierarchy of linguistic information, processing in both the MT and the speech technology domains can avail of structured information at various points in the hierarchy and thus utilise both top-down and bottom-up information (cf. Figure 1).

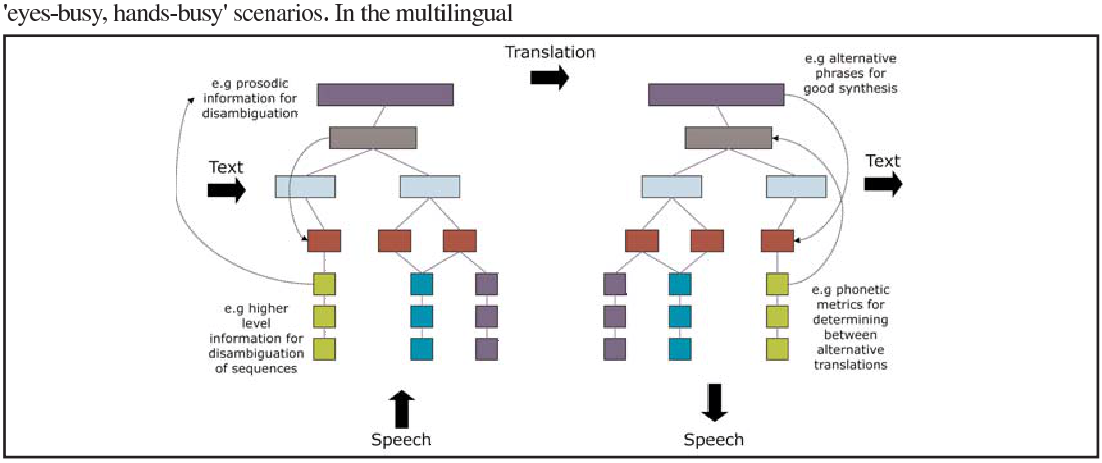

Figure 1. Potential benefits of integrated language technologies

Text-to-text MT may also use the speech resources where this information can support disambiguation. We envisage the MT and speech technology constraints working together to come up with the best solution, where gains outweigh any disadvantages in terms of added complexity. Building on our previous work in speech synthesis (Cahill & Carson-Berndsen 2006), and motivated by a concatenative approach to synthesis, the synthesiser is designed to learn automatically a new language from speech data and a pronunciation dictionary. This novel approach includes an adaptive, motivated cost function which uses phonetic insights and phonotactic and morphotactic information on pronunciation and speaker variation facilitating explicit integration with the MT engines described above for use in assistive automatic interpreting. Phonetic insights on voice characteristics are being used to develop a synthesis engine that models specific emotive state to endow synthetic speech with a more human, expressive character (Gobl & Ní Chasaide 2003).

For the ASR engine, we are building on the MuSE speech recognition system, developed by Kelly et al. (2007). This system combines flexible, robust feature-extraction engines with a syllable-recognition component based on language-specific finite-state phonotactic models of speech to overcome the problems of speaker and language dependence (Carson-Berndsen 2000, Aioanei et al. 2005). We are extending MuSE using ML techniques to develop feature-extraction engines to detect phonetic characteristics of speech that are relevant across many languages (Kanokphara et al. 2006). Phonetic feature-extraction engines produce multilinear representations of speech utterances and provide a way of modelling and investigating variability; feature-based inheritance hierarchies provide information on well-formed segments and deal with underspecification. Each of the ASR engines (feature, phoneme, syllable, word) will seek to integrate explicitly with, and utilise, higher-level linguistic information from MT and TA (cf. Figure 1).

The specific outcomes of this research will be an intelligent speech-synthesis engine integrated with the MT engines described above based on a novel unit-selection approach using hierarchies of linguistically motivated units together with phonetic insights to generate natural-sounding speech. We will also develop an intelligent ASR system, integrated with MT engines, consisting of a set of feature-extraction engines and a linguistic recognition model which takes the parallel feature streams as input and outputs orthographic word sequences and annotations required by MT and TA. The speech synthesis and speech recognition engines will enable speech interfaces and facilitate access and personalisation for eyes-busy, hands-busy localisation applications. The engines will be evaluated not only with respect to standard speech synthesis and recognition metrics but also in the context of demonstrator applications with MT and adaptive content.

4. Text Analytics

The key research challenges addressed in this area are the design, development and evaluation of multilingual text-type and genre classifiers for the purposes of localisation, the automation of localisation metadata annotation to support localisation workflows, and the automatic dependency annotation for syntax-enhanced SMT and EBMT engines, and novel probabilistic transfer-based MT systems.

Building on our previous work (O'Brien & Vogel 2003, Kelleher & Luz 2005, Davy & Luz 2007), we are designing, developing and evaluating a suite of multilingual ML-based text classifiers and exploring them in the context of generating localisation workflow metadata to support the automatic choice of MT engines fine-tuned to text-type and genre. Li & Vogel (2010) continue to refine classification methods. A further application of this work relates to improved semantics of user queries related to the digital content management research described by O'Connor et al. (2009). This task encompasses three distinct problems: automatic clustering, assessment of corpus homogeneity and assignment of category labels. Although these three problems have been widely studied in general contexts, applications to MT and localisation pose novel challenges in terms of scarcity of annotated data and automatic data gathering from Web sources. Active learning techniques (Davy & Luz 2007) can help tackle the first issue; measures sensitive to hyperlink structure (Kelleher & Luz 2005) can help select effective training data from online sources. In addition, building on our previous work (Cahill et al. 2004), we are scaling our multilingual dependency annotation technology to GigaWord and Web corpora to support the automatic acquisition of probabilistic transfer-based MT and syntax-enhanced EBMT and SMT engines.

Classification may focus on the level of the sentence

and its constituents or at larger intersentential levels such as the level of a document or corpus. Research continues in both directions and with due attention to multilingual considerations. At intrasentential levels, yet driven by multilingual corpus based needs, we have addressed syntactic analysis in LFG. Considerable activity at DCU has built upon a substantial platform of probabilistic parsing and generation technologies developed there and in concert with the LFG and MT communities (Bryl et al. 2009, Graham et al. 2009). A great wealth of linguistic resources in the form of treebanks have already emerged. On the semantic annotation side we have been addressing word-sense discrimination and labelling of the semantic roles that arguments fill with respect to predicates (Li et al. 2009). Language guessing is an established application of text classification techniques (Cavnar & Trenkle 1994) and in scenarios made possible by success within NGL research, multilingual chat for example, the role of language guessing for individual sentence-level contributions for the purposes of directing them to particular MT engines is possibly more clear than for high-volume document localization. However, in the industry context, automatic classification of document components on the basis of style and content, as relevant to the legal section versus hardware requirements section versus operational use sections is also relevant in the process of speeding texts on to appropriate translators. Our research into stylistic classification has use with respect to assessment of source-language conformity with house styles and general stylistic homogeneity, evaluation of translation outputs, identification of effects of source language and translator on translations, and the effects of interaction on language production. A range of scientific questions and practical applications are explored.This research has several aims:

(a) to develop and evaluate multilingual text classifiers for Next Generation Localisation;
(b) to develop and evaluate ML-based localisation workflow metadata annotators; and
(c) to develop and scale automatic multilingual dependency annotation technology to support automatic acquisition of novel transfer-based probabilistic MT and syntax-enhanced SMT and EBMT engines.

We seek to quantify theoretical bounds on the effectiveness of methods on the basis of internal properties of data under scrutiny. The research aims to exploit text classification methods in a range of scenarios that present purely theoretical problems and theoretical problems that have practical industrial relevance, not only those alluded to above.

5. Crowd-sourcing and integration

In close engagement with our industry partners, we have identified three significant emerging developments in the localisation landscape. These are:

(a) crowd-sourcing and community platforms;
(b) integration of MT systems and TMs; and
(c) interaction of MT and translation post-editing (both manual and automatic post-editing).

5.1. Crowd-sourcing

Over the last two years crowd-sourcing and the supporting community platforms have begun to take centre stage in localisation. Google, Facebook, IBM, Microsoft, Adobe, Symantec and Sun (amongst others) have successfully involved users (and customers) in a variety of localisation-related efforts, ranging from allowing users to correct the output of SMT systems and using the corrections to re-train the MT systems (Berlin 2009, Cohen 2009), to fully involving users in the localisation of web-pages and interfaces (Hosaka 2008). Crowd-sourcing is attractive as it can reduce the cost base associated with localisation activities (users do it for 'free'), and help tune the localised output to the (linguistic) requirements and expectations of a particular user (or 'fan') base. Facebook for example, having involved users to localise their social networking sites into French, German and Spanish, has reported an increase from 52M visitors to 124M as a result (Britton and McGonegal 2007:80; Eskelsen et al. 2008:120), while IBM report[3] that in the first year of its launch, 3,000 employees contributed 36M words' worth of translations. On the other hand, translators have complained that crowd-sourcing devalues their profession, when they are expected to work for free or for payment in kind (Newman 2009).

Important issues in crowd-sourcing are quality control and text domain. Facebook, for example, uses translator rankings on their sites to publicise and hence reward user-supplied translations, voting-based translation (to eliminate 'bad' translations) as well as professional translators to validate and post-edit user-localised sites before going live. Berlin (2009) mentions the need for "review by a second

3 http://www.research.ibm.com/social/projects_nfluent.html

translator before publication and [to] have translators sign their work, discouraging sloppy or deliberately malicious translations". It is said that users are only interested in translating customer-facing content (i.e. the main web-pages visible to a Facebook client) and (unsurprisingly) are not interested in translating technical or legal documentation pertaining to the sites, which are localised using fully professional localisation operations.

Crowd-sourcing can be used to provide essential

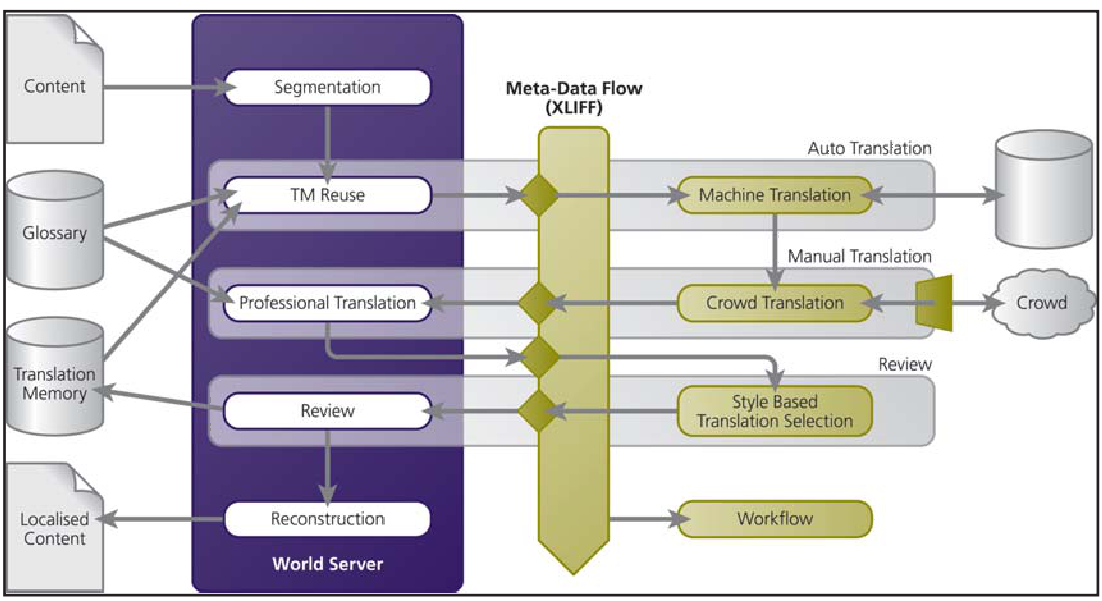

Figure 2. Localisation workflow integrating crowd-sourcing

localisation services to NGOs and Development Agencies where traditional ROI calculations do not apply. Our current research involves Translators Without Borders[4] (as a user base), an organisation which provides free localisation services to NGOs such as Médecins Sans Frontières and Ashoka. Crowd-sourcing is fully integrated with the technology developments relating to MT and speech technology (see Figure 2): with access to the MT systems for post-editing, the resulting corrections will be used to retrain the MT systems.

5.2. Translation Memories

TMs are a core technology in state-of-the-art localisation workflows (Schäler 1996). In a sense, a basic TM constitutes a (very simple) EBMT system which operates at the level of complete sentences (Somers & Fernández Díaz 2004). Usually TM systems are enhanced through partial match facilities, if a complete match cannot be found. In general, partial matches come in two forms: (i) partial matches on the sentence level and (ii) complete matches on subparts of the original sentence. The first supports retrieval of translations for sentences that are similar to (but not the same as) the original sentences and the second allows matches on parts of the input string and retrieves potentially useful translations for those fragments. In each case, the TM flags the translations retrieved from partial matches to the human translator, who will post-edit the output proposed by the TM. Leveraging TMs can substantially reduce translation costs and is particularly effective for predictable text types (including technical documentation, user manuals, help files etc.).

With increased availability and continuously improving output quality, MT is beginning to make strong inroads into localisation operations. In principle, TM and MT technologies are complementary: TMs provide maximal quality performance on seen material, while MT is likely to perform better on unseen data. Combining the complementary strengths of TMs and MT approaches in the localisation workflow promises efficiency and

4 http://tsf.eurotexte.fr/index-en.shtml

quality gains over and above the exclusive use of either technology in isolation. The challenge is provided by data in between, i.e. data (sentences) consisting of a mix of seen and unseen components. Depending on the text type, this type of data can be the most frequent. To date, the optimal combination between TM and MT technology is not known (cf. Simard and Isabelle 2009, for one recent view on this). In our research we will parameterise the problem according to a number of important (controlled) variables, including:

(a) MT type: rule-based vs. data-driven;
(b) text type: predictable vs. unpredictable;
(c) MT output quality: good vs. poor;
(d) language pairs: closely or distantly related and
(e) well- or poorly-resourced.

We expect that the space defined by these variables will lead to different optimal combinations of TM and MT technologies in the workflows. For example, MT in its example-based form can be seen as an extension of a TM in that while using a TM it is up to the human translator to decide what to do with the proposed match, in EBMT the system takes the match and tries to manipulate it to provide a translation. One way in which the performance of a TM can be improved is where a match differs minimally from the text to be translated: by connecting the TM to target-language resources such as parallel corpora or even simple dictionaries, the target-language text to be changed can be highlighted, and a translation proposed - a rudimentary type of EBMT. Another example is that with data-driven MT systems (such as SMT) we can automatically generate TMs from the training sections of the SMT system and use the resulting TM/MT combination in the space defined above; in the other direction, we can use the aligned text in TMs and train an SMT system.

5.3. Post-editing

Manually post-editing MT output is an emerging task in localisation workflows. However, to date, translators are rarely trained, if at all, to post-edit MT output - a quite different task from revising human translations (Loffler-Laurian 1985, McElhaney & Vasconcellos 1988:141, Allen 2003) - and computerised translation tools (such as editors in translators' workbenches) do not specifically support post-editing tasks. In fact, post-editing is often viewed highly unfavourably by professional translators. Our research investigates post-editing strategies in localisation scenarios, to develop recommendations for how a post-editing interface would support the observed strategies and identify training needs for translators using MT support. We are interested in measuring correlations, if they exist, between MT evaluation metrics (e.g. Translation Edit Rate, Snover et al. 2006) or MT system-generated confidence scores and post-editing effort, measured in terms of the time taken to complete the task or the number of edits made in the segment. We are also interested in correlations between post-editing effort and linguistic features such as length of the source segment, number of verbs in the source segment, number of nouns or noun phrases etc. Additionally, our research includes investigations into correlations between years of professional experience and post-editing effectiveness.

Automatic statistical post-editing is a new and exciting area in MT. In statistical post-editing, the output of one MT system is used to train a second-stage MT system, which operates on the output of the first MT system, with the intention to improve overall translation quality. To date, such system cascades have used a first-stage rule-based MT system followed by an SMT system (e.g. Simard et al. 2007). Improvements in automatic translation quality will, of course, correspond to cost savings in localisation workflows (due to reductions in post-editing efforts in off-line scenarios and the ability to post unedited quality output in on-line scenarios).
The central research idea is to generalise this architecture to a recursive target-side cascaded self-training architecture for MT, using a variety of ML-based MT architectures.

6. Conclusion

Basic research across ILT continues in collaboration with industry partners where mutual scientific interests also have industrial relevance. The research programme has been dynamic in responding to emerging problems and will continue to do so at the same time that foundational matters in text analytics are explored.

Acknowledgement

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University, Trinity College Dublin, and University College Dublin.

References

Aioanei, D., Neugebauer, M. and Carson-Berndsen, J. (2005) 'Validation techniques for parallel feature streams: the case of phoneme identification for speech recognition', Archives of Control Sciences, 15, 279-290.

Allen, J. (2003) 'Post-editing', in Somers, H., ed., Computers and Translation: A Translator's Guide, Amsterdam: John Benjamins, 297-317.

Bangalore, S. and Joshi, A. (1999) 'Supertagging: An approach to almost parsing', Computational Linguistics, 25, 237-265.

Berlin, L. (2009) 'Translating online content for love of language: Volunteers provide nuance not available with automated systems, International Herald Tribune, 18 May.

Britton, D.B. and McGonegal, S. (2007) The Digital Economy Fact Book, Ninth Edition 2007, Washington, D.C.: The Progress & Freedom Foundation.

Brown, R.D. (1999) 'Adding linguistic knowledge to a lexical example-based translation system', in Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99), Chester, England, 22-32.

Bryl, A., van Genabith, J. and Graham, Y. (2009) 'Guessing the grammatical function of a non-root f-structure in LFG', in 11th International Conference on Parsing Technologies (IWPT'09), Paris, France, 146-149.

Cahill, A., Burke, M., McCarthy, M., O'Donovan, R., van Genabith, J. and Way, A. (2004) 'Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations', in ACL-04, 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 319-326.

Cahill, P. and Carson-Berndsen, J. (2006) 'The Jess Blizzard Challenge 2006 entry', in Blizzard Challenge 2006 Workshop, Interspeech 2006 - ICSLP, Pittsburgh, PA., [4 pages]

Cahill, A. and van Genabith, J. (2006) 'Robust PCFG-based generation using automatically acquired LFG approximations', in COLINGoACL Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 1033-1040.

Callison-Burch, C., Osborne, M. and Koehn, P. (2006) 'Re-evaluating the role of BLEU in machine translation research', in EACL-2006, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 249-256.

Carl, M. and Way, A., eds. (2003) Recent Advances in Example-Based Machine Translation, Dordrecht: Kluwer.

Carson-Berndsen, J. (2000) 'Finite state models, event logics and statistics in speech recognition', in Spärck Jones, K.I.B., Gazdar, G.J.M. and Needham, R.M., eds., Computers, Language and Speech: Integrating formal theories and statistical data, Philosophical Transactions of the Royal Society, Series A, vol. 358 (1769), 1255-1266.

Carson-Berndsen, J. and Walsh, M. (2005) 'Phonetic time maps: Defining constraints for multilinear speech processing', in Barry, W.J. and van Dommelen, W., eds., The Integration of Phonetic Knowledge in Speech Technology, Dordrecht: Springer, 45-66.

Cavnar, W.B. and Trenkle, J.M. (1994) 'N-gram-based text categorization', in Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 161-175.

Chiang, D. (2005) 'A hierarchical phrase-based model for statistical machine translation', in 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, 263-270.

Clark, S. and Curran, J. (2004) 'The importance of supertagging for wide-coverage CCG parsing', in Coling-2004: 20th International Conference on Computational Linguistics, Geneva, Switzerland, 282-288.

Cohen, N. (2009) 'A translator tool with a human touch', The New York Times, 22 Nov.

Daelemans, W. and van den Bosch, A. (2005) Memory-Based Language Processing, Cambridge: Cambridge University Press.

Davy, M. and Luz, S. (2007) 'Active learning with history-based query selection for text categorisation', in Amati, G., Carpineto, C. and Romano, G., eds., Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, LNCS 4425, Dordrecht: Springer, 695-698.

Deng Y. and Byrne, W. (2005) 'HMM word and phrase alignment for statistical machine translation', in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, pp. 169-176.

Deng Y. and Byrne, W. (2006) 'MTTK: An alignment toolkit for statistical machine translation', in Proceedings of the Human Language Technology Conference of the NAACL, New York City, NY, 265-268.

Eskelsen, G., Marcus, A. and Ferree, W.K. (2008) The Digital Economy Fact Book, Tenth edition, 2008, Washington, D.C.: The Progress & Freedom Foundation.

Galron, D., Penkale, S., Way, A. and Melamed, I.D. (2009) 'Accuracy-based scoring for DOT: Towards direct error minimization for data-oriented translation', in EMNLP 2009, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 371-380.

Gobl, C. and Ní Chasaide, A. (2003) 'The role of voice quality in communicating emotion, mood and attitude', Speech Communication, 40, 189-212.

Graham, Y., Bryl, A. and van Genabith, J. (2009) 'F-structure transfer-based statistical machine trans-lation', in Proceedings of Lexical Functional Grammar 2009, 14th International LFG Conference, Cambridge, UK [2 pages].

Groves, D. (2007) Hybrid Data-Driven Models of Machine Translation, unpublished thesis (PhD), Dublin City University.

Groves, D., Hearne, M. and Way, A. (2004) 'Robust sub-sentential alignment of phrase-structure trees', in COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 1072-1078.

Groves, D. and Way, A. (2005) 'Hybrid data-driven models of MT', Machine Translation, 19, 301-323.

Groves, D. and Way, A. (2006) 'Hybridity in MT: Experiments on the Europarl corpus', in Proceedings of the 11th Conference of the European Association for Machine Translation, Oslo, Norway, 115-124.

Haque, R., Naskar, S., Ma, Y. and Way, A. (2009a) 'Using supertags as source language context in SMT', in Proceedings of EAMT-09, the 13th Annual Meeting of the European Association for Machine Translation, Barcelona, Spain, 234-241.

Haque, R., Naskar, S., van den Bosch, A. and Way, A. (2009b) 'Dependency relations as source context in phrase-based SMT', in Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, (forthcoming).

Haque, R., Naskar, S., van Genabith, J. and Way, A. (2009c) 'Experiments on domain adaptation for English-Hindi SMT', in Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, (forthcoming).

Hassan, H., Hearne, M., Way, A. and Sima'an, K. (2006) 'Syntactic phrase-based statistical machine translation', in IEEE/ACL 2006 Workshop on Spoken Language Translation, Palm Beach, Aruba, 238-241.

He, Y. and Way, A. (2009a) 'Learning labelled dependencies in machine translation evaluation', in Proceedings of EAMT-09, the 13th Annual Meeting of the European Association for Machine Translation, Barcelona, Spain, 44-51.

He, Y. and Way, A. (2009b) 'Improving the objective function in minimum error rate training', in Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada, 238-245.

Hearne, M. and Way, A. (2006) 'Disambiguation strategies for data-oriented translation', in 11th Annual Conference of the European Association for Machine Translation - Proceedings, Oslo, Norway, 59-68.

Hosaka, T.A. (2008) 'Facebook asks users for free translations of Web site's new international versions', AP Worldstream, 18 Apr.

Kanokphara, S., Macek, J. and Carson-Berndsen, J. (2006) 'Comparative study: HMM and SVM for automatic articulatory feature extraction', in Ali, M.

and Dapoigny, R., eds., Advances in Applied Artificial Intelligence: 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006, Annecy, France, LNCS 4031, Berlin: Springer, 674-681.

Kelleher, D. and Luz, S. (2005) 'Automatic hypertext keyphrase detection', in IJCAI-05, Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 1608-1609.

Kelly, R., Carson-Berndsen, J., Macek, J., Aioanei, D., Kanokphara, S. and Cahill, P. (2007) MuSE: The Muster Speech Engine, MUSTER Technical Report, School of Computer Science and Informatics, University College Dublin.

Koehn, P., Och, F.J. and Marcu, D. (2003) 'Statistical phrase-based translation', in Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Alberta, Canada, 127-133.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin A. and Herbst, E. (2007) 'Moses: Open source toolkit for statistical machine translation', in ACL 2007 Proceedings of the Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 177-180.

Lambert, P., Ma, Y., Ozdowska, S. and Way, A. (2009) 'Tracking relevant alignment characteristics for machine translation', in Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada, 268-275.

Li, B., Emms, M., Luz S. and Vogel, C. (2009) 'Exploring multilingual semantic role labeling', in Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task, Boulder, Colorado, 73-78.

Li, B. and Vogel, C. (2010) 'Leveraging sub-class partition information in binary classification and its application', in Bramer, M., Ellis, R. and Petridis, M., eds., Research and Development in Intelligent Systems XXVI, Incorporating Applications and Innovations in Intelligent Systems XVII, London: Springer, 299-304.

Loffler-Laurian, A.-M. (1985) 'Traduction automatique et style', Babel, 31 (2), 70-76.

Ma, Y., Stroppa, N. and Way, A. (2007) 'Bootstrapping word alignment via word packing', in ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 304-311.

Ma, Y. and Way, A. (2009) 'Bilingually motivated domain-adapted word segmentation for statistical machine translation', in EACL 2009, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 549-557.

Magerman, D.M. (1995) 'Statistical decision-tree models for parsing', in 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, 276-283.

Marcu, D. and Wong, W. (2002) 'A phrase based, joint probability model for statistical machine translation', in Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, 133-139.

McElhaney, T. and Vasconcellos, M. (1988), 'The translator and the postediting experience', in Vasconcellos, M., ed., Technology as Translation Strategy, Binghamton, NY: State University of New York at Binghamton (SUNY), 140-148.

Nagao, M. (1984) 'A framework of a mechanical translation between Japanese and English by analogy principle', in Elithorn, A. and Banerji, R., eds., Artificial and Human Intelligence: Edited Review Papers at the International NATO Symposium on Artificial and Human Intelligence Sponsored by the Special Programme Panel Held in Lyon, France October, 1981, Amsterdam, North-Holland: Elsevier Science Publishers, 173-180.

Newman, A.A. (2009) 'Translators wanted for LinkedIn, especially if they don't ask for any pay: Does Web site's request exploit professionals or give them exposure?', International Herald Tribune, 30 Jun.

O'Brien, C. and Vogel, C. (2003) 'Spam filters: Bayes vs. chi-squared; letters vs. words', in Proceedings of the 1st International Symposium on Information and Communication Technologies, Dublin, Ireland, 298-303.

O'Brien, S. (2003) 'Controlling Controlled English: An analysis of several controlled language rule sets', in Controlled Language Translation, EAMT-CLAW-03, Dublin, Ireland, 105-114.

O'Brien, S. and Roturier, J. (2007) 'How portable are controlled language rules? A comparison of two empirical MT studies', in MT Summit XI, Copenhagen, Denmark, 345-352.

Och, F.J. (2003) 'Minimum error rate training in statistical machine translation', in 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 160-167.

Och, F.J. and Ney, H. (2002) 'Discriminative training and maximum entropy models for statistical machine translation', in ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, 295-302.

Och, F.J. and Ney, H. (2003) 'A systematic comparison of various statistical alignment models', Computational Linguistics, 29, 19-51.

Owczarzak, K., Groves, D., van Genabith, J. and Way, A. (2006) 'Contextual bitext-derived paraphrases in automatic MT evaluation', in HLT-NAACL 06 Statistical Machine Translation, Proceedings of the Workshop, New York City, USA, 86-93.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002) 'BLEU: A method for automatic evaluation of machine translation', in ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, 311-318.

Schäler, R. (1996) 'Machine translation, translation memories and the phrasal lexicon: The localisation perspective', in EAMT Workshop TKE '96, Vienna, Austria, 21-34.

Simard, M., Goutte, C. and Isabelle, P. (2007) 'Statistical phrase-based post-editing', in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, 508-515.

Simard, M. and Isabelle, P. (2009) 'Phrase-based machine translation in a computer-assisted translation environment', in MT Summit XII: Proceedings of the Twelfth Machine Translation Summit, Ottawa, ON, Canada, 120-127.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006) 'A study of translation edit rate with targeted human annotation', in AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation", Cambridge, MA, 223-231.

Somers, H. (1999) 'Review article: Example-based machine translation', Machine Translation, 14, 113-158; repr. (revised) as 'An overview of EBMT' in Carl, M. and Way, A., eds., Recent Advances in Example-Based Machine Translation, Dordrecht (2003): Kluwer, 3-57.

Somers, H. and Fernández Díaz, G. (2004) 'Translation memory vs. example-based MT: What is the difference?', International Journal of Translation, 16 (2), 5-33.

Srivastava, A. and Way, A. (2009) 'Using percolated dependencies for phrase extraction in SMT', in Proceedings of the Twelfth Machine Translation Summit, Ottawa, Canada, 316-232.

Stroppa, N., van den Bosch, A. and Way, A. (2007) 'Exploiting source similarity for SMT using context-informed features', in TMI-2007: Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation, Skövde, [Sweden], 231-240.

Tinsley, J., Ma, Y., Ozdowska, S. and Way, A. (2008) 'MaTrEx: The DCU MT System for WMT 2008', in ACL-08: HLT, Third Workshop on Statistical Machine Translation, Columbus, Ohio, 171-174.

Tinsley, J. and Way, A. (2009) 'Parallel treebanks and their exploitability in machine translation', Machine Translation (in press).

Ueffing, N., Haffari, G. and Sarkar, A. (2007) 'Transductive learning for statistical machine translation', in ACL 2007: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 25-32.

van den Bosch, A., Stroppa, N. and Way, A. (2007) 'A memory-based classification approach to marker-based EBMT', in METIS-II Workshop: New Approaches to Machine Translation, Leuven, Belgium, [10 pages]

van Genabith, J. (2009) 'Next generation localisation', Localisation Focus, The International Journal of Localisation, this volume

Way, A. and Gough, N. (2003) 'wEBMT: Developing and validating an EBMT system using the World Wide Web', Computational Linguistics, 29, 421-457.

Way, A. and Gough, N. (2005a) 'Comparing example-based and statistical machine translation', Journal of Natural Language Engineering, 11, 295-309.

Way, A. and Gough, N. (2005b) 'Controlled translation in EBMT', Machine Translation, 19, 1-36.
O'Connor, A., Lawless, S., Zhou, D., Jones, G. J. F, Way, A. (2009) 'Applying Digital Content Management to Support Localisation', Localisation Focus, The International Journal of Localisation, this volume