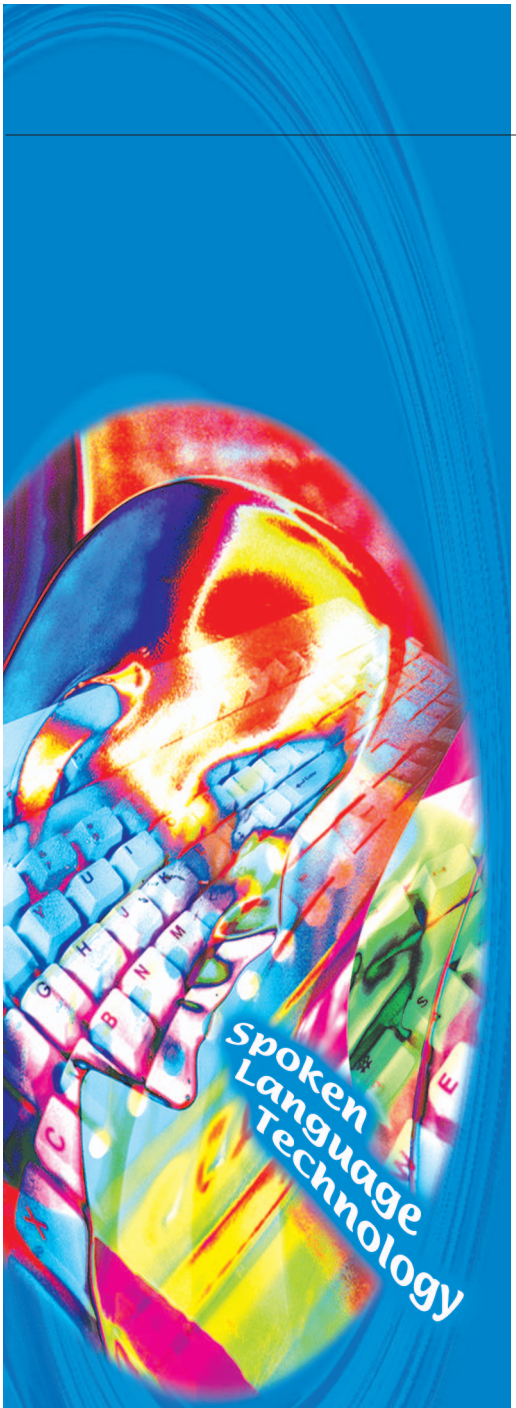

Pascale Fung and Tanja Schultz

Multilingual Spoken Language Processing

Challenges for multilingual systems

In the past decade, the performance of spoken language understanding systems has improved dramatically, including speech recognition, dialog systems, speech summarization, and text and speech translation. This has resulted in an increasingly widespread use of speech and language technologies in a wide variety of applications. With more than 6,900 languages in the world and the current trend of globalization, one of the most important challenges in spoken language technologies today is the need to support multiple input and output languages, especially if applications are intended for international markets, linguistically diverse user communities, and nonnative speakers. In many cases these applications have to support even multiple languages simultaneously to meet the needs of a multicultural society. Consequently, new algorithms and tools are required that support the simultaneous recognition of mixed-language input, the summarization of multilingual text and spoken documents, the generation of output in the appropriate language, or the accurate translation from one language to another. This article surveys significant ongoing research programs as well as trends, prognoses, and open research issues with a special emphasis on multilingual speech processing as described in detail in [26] and multilingual language processing as presented in [6].

CHALLENGES AND SOLUTIONS FOR MULTILINGUAL TECHNOLOGIES AND APPLICATIONS

LACK OF DATA RESOURCES

For more than 20 years speech recognition has been dominated by statistically based modeling schemes, such as hidden Markov models (HMMs) and n-gram language models, and despite their limits both schemes have turned out to be very successful. Also, in recent years the paradigms of speech and text translation noticeably shifted from Interlingua-based or example-based toward statistical methods. The core algorithms applied in statistical modeling are primarily language independent and have

Digital Object Identifier 10.1109/MSP.2008.918417

1053-5888/08/$25.00©2008IEEE

indeed proven to work reasonably well for a variety of languages or language pairs. On the flip side, statistical models heavily rely on vast amounts of training data for robust and reliable parameter estimation. For high-end performance, thousands of hours of (quickly) transcribed audio data and hundreds of millions of words of monolingual text data or millions of words of bilingual text corpora are typically used.

As the public demand turns toward less widespread languages, it becomes clear that this traditional approach is prohibitive to all but the most widely spoken, widely read, and economically viable languages. The Linguistic Data Consortium has managed the design and collection of large databases for languages of interest, and the European Language Resources Association also provides databases in multiple languages with an emphasis on European languages. Nevertheless, large-scale data resources have been systematically collected and distributed for less than 50 languages. Furthermore, very few data collections emphasize uniform acquisition scenarios across languages, which are crucial for truly multilingual systems. One of the few exceptions is GlobalPhone, a standardized multilingual text and speech database [27], which provides transcribed speech data for the development and evaluation of multilingual spoken language processing systems in the most widespread languages of the world. GlobalPhone is designed to be uniform across languages with respect to the amount of text and speech data per language, the audio quality, the collection scenario, and the transcription conventions. As a consequence, GlobalPhone supplies an excellent basis for research in the areas of 1) multilingual speech recognition and synthesis, 2) rapid deployment of speech processing systems to new languages, 3) language and speaker identification tasks, and 4) monolingual speech recognition and synthesis in a variety of languages. To date, GlobalPhone covers 18 languages, including Arabic, Bulgarian, CH-Mandarin, CH-Shanghai, Croatian, Czech, French, German, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, and Turkish. In total the corpus contains 350 hours of speech spoken by 1,700 native adult speakers.

ONE OF THE CENTRAL ISSUES IN BUILDING SPOKEN LANGUAGE SYSTEMS FOR NEW LANGUAGES IS TO BRIDGE THE GAP BETWEEN LANGUAGE AND TECHNOLOGY EXPERTISE.

In response to the small number of languages with an appropriate amount of text resources compared to the huge number of living languages, various strategies and algorithms have been proposed to mine monolingual and parallel text data. In [16] the authors first identify a resource-rich language very similar to the target language and then extract useful statistics from the resource-rich data to project them back into the target language. For translation the focus is on mining parallel data from monolingual data since statistical machine translation systems are trained from parallel, mostly human-translated bilingual data, which are hard to come by. Such data are typically harvested from governmental and political documents, and more recently from the Web. Fung and Cheung [7] and others have devised algorithms that bootstrap from comparable bilingual texts and then extract parallel sentences. The mining systems are trained from parallel data to know what constitutes as a pair of translated sentences. After this training phase they search for relevant documents, i.e., contemporaneous articles, and those that are likely to contain parallel sentences of similar style.

LACK OF TOOLS AND LINGUISTIC RESOURCES

The second major bottleneck to system development is the lack of linguistically specific resources and tools crucial to all spoken language applications, such as pronunciation dictionaries, bilingual dictionaries, lexicons or phrase tables, word segmenters for writing systems that do not provide natural word segmentation, part-of-speech taggers, as well as syntactic and semantic parsers. Although statistical algorithms can be applied to automatically train such tools, it requires some annotated data, which is even harder and costlier to come by than any unannotated data.

Several language-independent methods have been proposed to automatically augment or construct pronunciation dictionaries [4] and colloquial or bilingual dictionaries [2] or to bypass the need of pronunciation dictionaries by multilingual grapheme-based approaches [14], [17]. The first two methods involve bootstrapping from existing mono- or bilingual pronunciation dictionaries and then building entries for unseen data from monolingual or from nonparallel data in multiple languages. The latter two methods are based on the letter-to-sound relationship between the written representation of a word and its spoken form, and its performance correlates with the closeness of the letter-to-sound relationship.

Another approach to solving the lack of linguistic tools is by coercing or projecting tools developed in a popular language into an under-resource language. Algorithms have been proposed to cross-lingually construct acoustic models for speech recognition [27] and speech synthesis [1], language models for speech recognition and translation [30], segmenters and part-of-speech taggers, or syntactic and semantic parsers [6]. More details about these approaches will be provided in the "Multilingual Technologies and Applications" section.

LACK OF LANGUAGE AND TECHNICAL EXPERTISE

With a suitable amount of data and well-defined standards, the development of speech processing components might be considered to be a rather straightforward task. However, this task turns out to be surprisingly time and cost intensive, which is partly due to language-specific peculiarities, but another reason is the challenge of finding language experts, especially for low-density languages [26]. Without a skilled language expert, speech system developers face the time-consuming situation of having to either familiarize themselves with the language in question or train an unskilled language

expert in speech and language technologies. Social and cultural aspects may further complicate the process, e.g., respect or politeness may prevent native users from identifying system flaws or may make communication between native informants and nonnative developers difficult. Consequently, one of the central issues in building spoken language systems for new languages is to bridge the gap between language and technology expertise [26].

After an overview of the state-of-the-art in multilingual spoken language technologies and applications in the next section, we will discuss proposed solutions to tackle the challenges of the lack of resources and to bridge the gap between language and technology expertise.

MULTILINGUAL TECHNOLOGIES AND APPLICATIONS

This section surveys state-of-the-art solutions to automatic learning and cross-lingual projection methods to develop resources and tools in a new language as well as to rapidly build multilingual acoustic and language models with little or no training data.

MULTILINGUAL AND MULTI-ACCENT ACOUSTIC MODELING

Over the course of the last decade extensive research has been done in building multilingual speech recognizers bootstrapped from language-independent acoustic models [24], [27]. Here, acoustic models for new target languages are estimated by borrowing data from various source languages for which such data are more plentiful, while using only very limited amounts of adaptation data from the target language itself. Language adaptation of acoustic models is performed in three main steps: 1) a language-independent, universal sound set is defined by applying either purely data-driven methods or building upon the definitions of similar sounds across languages as documented in international phonetic alphabets like IPA, or a combination of these two methods; 2) multilingual, language-independent acoustic models are trained for the universal sound classes by sharing data across multiple languages; and 3) these models are then cross adapted to new target languages using very limited language-specific data. These steps require extensive experience, access to multilingual text and speech data, and a collection of monolingual speech recognition engines in multiple languages. Much of this research has been performed on the above-mentioned GlobalPhone data set [27] as well as in [24]. Cross-language adaptation has shown to be very useful particularly when data of the target language are small. Here the adaptation of multilingual decision trees to the new target language gave very promising results [27]. Furthermore, a combination of knowledge and data-driven approaches for the determination of acoustic-phonetic unit similarities seems to be beneficial. Language-independent acoustic models have also shown to be successful in the presence of nonnative speech recognition [32]. Recently, similar procedures have been applied to cross-dialectal recognizers using heterogeneous data [18].

Automatic recognition of spontaneous speech often needs to handle accents, i.e., speech influenced by the speaker's first language on the target language, which can happen on the acoustic or phonological level but might also impact the sentence construction. Liu and Fung [23] studied the problem of multiple regional accents in Mandarin Chinese (Putonghua) for a speaker-independent and accent-independent speech recognition task. Since most speakers acquired Putonghua as a second language, their pronunciations are strongly influenced by their native regional language. As a result, Putonghua-trained automatic speech recognition (ASR) systems perform poorly for nonnative accented speech. Conventional methods such as phone set extension to include accent-specific units are a common way to model phonetic changes. However, the extended phone set and augmented pronunciations may introduce more lexical confusion in the decoder. Acoustic model parameters are commonly modified to model acoustic changes in accented speech and approaches include full retraining of the acoustic model using a large amount of accented speech [12], applying maximum a posteriori (MAP) or maximum log likelihood ratio (MLLR) adaptation to fit the characteristics of a particular accent [26, Chapter 9], and using discriminative training to refine acoustic models [15]. A major weakness is that the parameters of acoustic models undergo an irreversible change, and the models lose their ability to cover other accents.

In [23], the authors propose a method to handle multiple regional accents together with standard speech by using different sets of accent-specific units with acoustic model reconstruction. Diversity of accent changes in multiple, accented speech is represented by different sets of accent-specific units. The results are very promising, showing ASR performances increases for both accented and standard Putonghua in the same system.

In summary, multilingual acoustic models show improvements over monolingual models, especially for rapid cross-language adaptation. In addition, multilingual acoustic models allow for the construction of truly multilingual systems, which can handle code switching and cross-lingual pronunciation effects. Last but not least, multilingual acoustic models are more compact, easier to maintain, and also important for applications on small footprint devices such as cell phones.

MULTILINGUAL PRONUNCIATION DICTIONARIES

For most automatic speech transcription systems, multilingual pronunciation dictionaries are still collections of monolingual dictionaries, which easily contain over 500,000 words per language in modern open-domain speech processing systems. These words can be divided into five types according to their function and corpus frequency: foreign and local proper names (millions each), technical content words (thousands), general content words (thousands), and function words (hundreds). While the latter two types of words are strongly language specific, the much more frequent technical words and proper names are often shared across languages. In [26, Chapter 5], Adda-Decker and Lamel investigated the number of common word entries for French, Spanish, German, and English and observed

that these languages share 20% of the most frequent 50,000 words in broadcast news, and that the proportion of imported words increases with vocabulary size. Proportions are particularly high for languages of small countries, which more easily incorporate words from other languages.

Lexical development involves various steps: 1) the choice of lexical units, 2) the selection of a vocabulary, and 3) the generation of word pronunciations. Normalization procedures are required to ensure viable lexical coverage, including the treatment of punctuation, numbers, abbreviations, acronyms, and capitalization. In the context of multilingual dictionaries the normalization of proper names is particularly important, as their transformation from the writing system of the origin into the target writing system can vary substantially due to the different transfer concepts; i.e., transliteration or transcription.

AUTOMATIC RECOGNITION OF SPONTANEOUS SPEECH OFTEN NEEDS TO HANDLE ACCENTS; I.E., SPEECH INFLUENCED BY THE SPEAKER'S FIRST LANGUAGE ON THE TARGET LANGUAGE.

In pronunciation development, phone sets need to compromise between pronunciation accuracy and modeling constraints. In a monolingual setup, rare phones often have to be eliminated since they do not occur frequent enough in the training corpus to allow for reliable model estimation. A multilingual setup may at least partially solve this problem as the chances are higher to see them more frequently. Furthermore, the richer phone set may support nonnative speech more properly. However, a finer distinction in the phone set requires more human effort and probably tight collaboration among various language experts to ensure completeness and consistency.

Multilingual dictionaries have been investigated particularly for small domains and applications, where many proper names are typically used, such as in car navigation systems or flight information systems. In the future we are likely to use word lists of millions of entries. A significant part of the vocabulary can then be shared among languages, while language-specific lexical entries can be limited to some tens of thousands of items. These aspects have not yet been addressed but offer new perspectives to multilingual spoken language processing.

MULTILINGUAL LANGUAGE MODELING

Many approaches have been proposed to tailor language models toward particular domains by text selection or various interpolation schemes. But what if a language model has to be created for languages where only few data resources are available? What if time and cost limitations require a rapid deployment? One promising approach is cross-lingual language model adaptation as proposed by [16]. The algorithm first identifies text data in a resource-rich language similar to the target language, then extracts useful statistics from those text corpora, and projects the statistics back into the target language. Another approach applicable to small domains is the use of grammar-based recognizers. Results on multilingual language modeling for multilingual speech interfaces indicate that some text-based knowledge such as named entities, content words, and phrases might be sharable across languages [5], while other parts of the grammar are not shareable but need to be trained from data [2]. Statistical methods are used to learn named entities, content words, and phrases from large amounts of a database, while a Markov process is proposed to learn the grammar automatically in [2]. Using multilingual language models and grammars is therefore one way to transfer knowledge across languages and also to perform integrated language identification [5]. In [2], the authors successfully created a language model for Cantonese using a small corpus of the resource-poor target language for interpolation with a large corpus of the resource-rich Mandarin language. Finally, grammars and statistical language models could also be intertwined to rapidly bootstrap larger domains from knowledge on smaller domains.

MULTILINGUAL SPEECH SYNTHESIS

As of today, speech recognizers in new languages are typically built by collecting several hours of well-recorded speech in the target language (Black and Lenzo, http://festvox.org). An alternative method is to apply the same idea as in speech recognition; namely, to use a multilingual acoustic model from an existing synthesizer in one language and cross adapt models to the target language based on a very small set of collected sentences [1], [20]. In order to utilize data and models from multiple speakers and multiple languages for speech synthesis, the existing methods of unit selection are unlikely to succeed given the wider variety of data and less consistency within it. However, with the work on HMM generation synthesis [31], we observe a move away from high-quality instance selection toward techniques that combine multiple instances into a parametric model. Within this paradigm, the selection of appropriate speakers via voice clustering techniques based on the HMM modeling approach allows for the combination of multiple speakers and even languages into single models and thus to build multilingual synthesizers. Latorre [20] and Black and Schultz [1] proposed to build such multilingual synthesizers using combined data from multiple languages. Black and Schultz [1] used the CLUSTERGEN synthesizer included in the FestVox suite, which offers a clustering technique for HMM-sizes segments. This is easy to implement within the framework of SPICE [29] (see also the "Bridging the Gap between Language and Technology Expertise" section) since both components, the recognizer and the synthesizer, share the same phone set. The CLUSTERGEN method depends on a reversible parameterization of speech. Feature vectors are clustered using classification and regression trees with features derived from the IPA phoneme scheme. At synthesis time the desired phones are then generated in the same way as for other synthesis techniques.

Both of the above-referenced studies showed that multilingual synthesis and cross-lingual adaptation are indeed feasible at

reasonable quality. However, a careful speaker selection for model combination is crucial to generate an optimal voice. Therefore, Black and Schultz [1] proposed speaker clustering techniques to optimize the model selection and showed that automatic speaker selection outperforms a manual selection. In summary, current studies indicate that factors such as lack of data, accuracy of data alignments, heterogeneity of speaker sets, and the complexity of multilingual phone inventories still pose major challenges to multilingual speech synthesis.

MULTILINGUAL DIALOG MODELING

In the past decade, we have seen the successful commercialization of spoken language systems all over the world. Perhaps the largest market share is taken by spoken dialog systems that serve to lessen the burden of human operators in call centers, to provide everyday information to travelers, and to enable easy access to online information by a simple phone call. Whereas the majority of these systems are in English, serving the North American market, the proliferation of such systems to the rest of the world with its higher population remains the objective of many system developers. Market acceptance of such systems is often hampered by cultural obstacles, as mentioned earlier. Whereas these issues remain to be solved by manual means, systems that are built with multilingual application in their design are poised to take advantage of this global market. As most commercial dialog systems are grammar based, it is quite easy to separate the language-specific component from the core speech engine. The separation of the application development module, which includes call flow and prompt design, enables core technology developers to collaborate with application developers, and each party is able to focus on its core competence. The driving force and beneficiaries of such modularization of core engine and application interface are the end users. Statistical models designed to learn dialog flow automatically from real-world data [21] are potential solutions for enabling more rapid development of multilingual dialog systems. Levin and Pierraccini [21] proposed a stochastic model based on Markov decision process for dialog systems. Their framework enables automatic training and adaptation of dialog systems and the objective evaluation of their performance. They recast the problem of dialog strategy design as an optimization problem that can be solved by different methods, such as reinforcement learning.

MULTILINGUAL SPEECH SUMMARIZATION

Spoken document summarization is the recognition, distillation, and presentation of spoken documents in a structural (and mostly) textual form. Applications of speech summarization include broadcast news extraction, meeting minute generation, and lecture speech translation.

Multilingual text and speech summarization has to deal with different layers of information variances caused by language differences at the signal, symbolic, lexical, semantic, and pragmatic levels. Speech recognition functions at the signal and symbol level, converting speech into text. Cross-lingual processing handles the lexical and semantic differences between the source and target languages. Ultimately, multilingual summarization also needs to deal with the pragmatic level differences between cultures represented by the source and target languages. There are differences in opinions and styles across languages even when they are describing the same event. Speech summarization handles the different layers of information in a feature-based approach. Acoustic, prosodic, and phonetic features are directly extracted from speech, and parameter strings and lexical, syntactic, and semantic features are extracted from the transcription.

Speech summarization is motivated by the fact that the distilled forms of spoken documents are often more useful to the audience. It poses additional challenges compared to text summarization. Speech contains various degrees of disfluencies and speaker variability. Speech recognition introduces errors. Moreover, spoken documents, unlike text documents, lack discernable structures such as titles, paragraph and sentence delimiters, punctuation, etc., to help with the decoding of the underlying features. This makes it difficult to apply text summarization methods directly to spoken documents. On the other hand, spoken documents have additional information that text documents lack, such as acoustic and prosodic information, often associated with the content of the speech. Speaker turns and dialog changes are also indicative of content switch.

Various research on Mandarin Chinese, English, and Japanese spoken document summarization has used more or less the same set of features. Results differ on the contribution of different features to the final summarization performance. Such differences, however, seem to be rather genre dependent than language-dependent. For both English [25] and Mandarin [35] it has been found that lexical features are not as important as acoustic and discourse features for broadcast news summarization. This is likely due to the style of news broadcasts and reports, which tend to follow a certain editorial style. News anchor speakers are professionally trained individuals who are apt at using prosodic features and tones to emphasize important points. This places a less stringent demand on recognition performance in speech summarization systems for broadcast news. On the other hand, for lecture speech and meeting speech, where the speakers tend to talk in a more diverse manner and with different degrees of spontaneity, acoustic and discourse features alone are no longer adequate. The most important features seem to be lexical and rhetorical, which are language dependent and require multilingual speech front ends. When incorporating these features most spoken document summarization systems employ an extractive approach, in which salient sentences or segments of speech are extracted and compiled into a final summary [11], [25].

The component that needs to extract a meaningful message from a transcript of garbled and disfluent speech is called "consolidation," as it involves removal of disfluencies, the combination of fragments across multiple turns, and the

extraction and summarization of key content. Such consolidation can also be realized by a statistical source-channel model acting as a summarizer that is extended to extract clean unambiguous sentences from garbled speech transcripts [10]. The module eliminates spontaneous speech phenomena (e.g., noise, disfluencies, false starts, repetitions) and attempts to find simpler, more efficient, more complete, and less ambiguous paraphrases of the input. It "translates" a corrupted, noisy input utterance into what might have been the intended well-formed sentence. This consolidation can be realized by either applying the translation engine as a front end to translate the document first, or by summarizing first and then translating the summary. However, since translating the entire garbled and disfluent speech output into the target language first would introduce more errors due to sparseness of appropriate training data, it is probably less error-prone to identify the salient segments before translation. Honal and Schultz [10] showed that the statistical approach can be adapted to match speaker-specific disfluency patterns and can also be efficiently ported to other languages. The models built on English spontaneous scheduling data were ported to Chinese conversational dialogs and showed very similar performance with minor parameter tuning.

Speech summarization relies to a large extent on the structural features in a voice document. These structural features are part of the discourse model. Documents in different languages tend to have different discourse models. Structural information of broadcast news might be similar across different languages as news reporters and anchor speakers are trained to use the same speaking style. On the other hand, spoken language style is still speaker dependent, and more significantly, culture/language dependent. Lectures and articles penned in different languages often have different rhetorical structure [8], making multilingual summarization and Q&A tasks more challenging.

Automatic methods have been proposed to extract discourse models in different languages. Fung and Ngai [8] presented a multidocument, multilingual, theme-based summarization system based on modeling text cohesion (story flow). They argue that inherent text cohesion exists and that it is specific to a particular story and specific to a particular language. Documents within the same story, and in the same language, share a common story flow, and this flow differs across stories and across languages. HMMs are proposed as story models and compared within and across stories as well as within and across languages (English and Chinese). The experimental results support the "one story one flow" and "one language one flow" hypotheses. This work holds hope in rendering the task of multilingual speech summarization less difficult as it automates the learning of language-specific discourse structures.

MULTILINGUAL SPEECH TRANSLATION

Speech translation is the transformation of a source language to a target language. Multilingual speech translation, meanwhile, must use a framework that can handle multiple pairs of source and target languages. In this section, we discuss challenges for speech translation on the one hand and those for multilingual speech translation on the other hand.

Speech translation has seen tremendous progress over the last couple of years. However, major problems remain that are very challenging. The first and foremost problem is the quality of automatic translation both in terms of adequacy and fluency. And since speech translation is typically applied to spontaneous speech, the recognition of the latter also introduces errors. Recognition errors and out-of-vocabulary words are particularly challenging to speech translation. The lack of punctuation in spoken language makes speech more difficult than text translation.

To handle the challenge of ambiguity introduced by spontaneous speech, most systems introduce semantic constraints by limiting the domain of discourse, thereby reducing the number of suitable interpretations. For many applications, constraining the domain is quite acceptable and can provide practical translation devices. Some statistical machine translation (SMT) systems [28], [36] focus on rapid development of the system in a new language, with small in-domain data and related domain data in addition to large out of domain data. Several existing systems reported performance gain from using ASR n-best lists, confusion networks, and by using word posterior probabilities from ASR. Some systems use sentence segmentation and punctuation prediction on the ASR output, thereby enhancing the performance of the subsequent translation system. Out-of-vocabulary words are handled variously by rule-based systems, task-related named entity lists, or by approximated replacement words in a morphologically rich language.

Optimal coupling of speech recognition and machine translation is another challenge. It is believed that a tighter integration of the ASR and machine translation modules helps mitigate some of the problems in both. Translation based on the recognition output lattice and using ASR scores in the latter stage of translation are two proposed methods for such purposes. Another problem is the cost of adding more languages to the translation system. As the number of languages of interest grows, the number of language pairs increases quadratically. State-of-the-art speech translation systems are dominated by the SMT approach today. For such systems, the challenge is a central model that is capable of handling the multitude of variations in word order, lexical units, morphology, and even semantic roles in different language pairs. Structure-based SMT approaches [33], [34], [9] are powerful in modeling the tree relations between languages pairs, thereby accounting for word-order differences. The inversion transduction grammar (ITG) approach [33] is particularly powerful in handling language pairs with vastly different word orders, such as Japanese and English.

The SMT approach uses automatic models that are trained on large, sometimes annotated bilingual corpora, which gives rise to the problem of the huge costs and time spent to develop resources for new language pairs. Tools have to be developed for morphological analysis, segmentation, chunking

and parsing, named entity tagging, etc., for each new language. Segmentation errors lead to translation errors. For some languages, such as Chinese, different encoding schemes give rise to segmentation errors. Meanwhile, in order to achieve good performance and coverage in large or even open domains, significant data resources have to be acquired for each new language. While large amounts of data in appropriate domains are hard to come by for the monolingual cases, it is even harder for bi- or multilingual purposes. We have discussed various approaches to mining bilingual and multilingual databases earlier in this article.

MULTILINGUAL ACOUSTIC MODELS ALLOW FOR THE CONSTRUCTION OF TRULY MULTILINGUAL SYSTEMS, WHICH CAN HANDLE CODE SWITCHING AND CROSS-LINGUAL PRONUNCIATION EFFECTS.

Another central technical challenge is the evaluation of translation performance. Measures such as BLEU and translation error rate (TER) can be applied to automatic evaluation, but both give only a rough estimate of the human translation error rate (HTER). More recently, metrics such as METEOR attempt to make up for the weaknesses of BLEU and TER by taking into account word classes and synonyms, as well as recall from multiple reference translations, in order to improve the correlation of automatic evaluation metrics with the performance of human interpreters. However, for this purpose manually or automatically generated semantic information needs to be acquired from multilingual resources such as Wordnet, FrameNet, and Hownet [6].

Last but not least, it is very difficult to give appropriate feedback to the user of a speech translation system. While users can easily judge a speech recognizer or synthesizer in their native language by simply checking the hypothesized output in form of written (or spoken) feedback, translation cannot be assessed unless the user is bilingual. Of course, a speaker who speaks both source and target language will not need a translation system. Consequently, speech translation systems need to give feedback in multiple languages. While Interlingua-based translation systems could use the Interlingua-based paraphrase as feedback, statistical-based translation systems usually retract to translate back from the target to the source language. This way the user can compare the original spoken utterance with the back-translated hypothesis. This comes at the costs of building two-way translation and has the risk of accumulating errors during the back-and-forth translation.

Over the years, research groups at Carnegie Mellon University and Hong Kong University of Science and Technology, among others, have developed a large number of speech translation systems for many languages, domains, and platforms. These efforts have shown that acceptable performance can be obtained for spontaneous speech input. Practical concerns such as portability and reconfigurability are also of increasing importance [28]. A noteworthy case is the commercial deployment of Japanese-to-English speech translation by the largest Japanese wireless service provider NTT DoCoMo via GSM, 3G, and the Japanese cell phone network. This translation system, designed and implemented by Advanced Telecommunications Research Institute International (ATR) in Kyoto, allows Japanese travelers to ask directions, book hotel rooms, and order food and drinks in English on a client-server platform. The system includes 1 million sentences of parallel corpora in the travel domain, speech corpora containing local accents, cross-lingual pronunciation dictionary, nonstationary noise suppression, minimum description length (MDL)-based HMM acoustic modeling, multiclass n-gram language modeling, phrase-based statistical machine translation, and a speech recognition-based distributed client-server system.

The two main challenges for portable speech translation systems are in maintaining the existing system and obtaining enough data for each new language and domain. These issues raise the questions of how to contain the cost of data collections and programming effort, which translation approaches to use, how to move from desktop prototypes to portable devices with limited computational power and memory footprint, how to integrate such portable devices, and finally how to deliver translation capabilities effectively.

MIXED-LANGUAGE PROCESSING

In the last 50 years, increasing levels of education, migration, mass communication, modernization, and globalization have led to the permeation of multilingualism, in general, and code switching, in particular. The term code switching refers to the usage of two or more languages in the same conversation by bilingual or multilingual speakers, typically resulting in mixed-language queries. The use of mixed language is more prevalent among bilingual cultures and especially in the technology and business communities. Many linguistics theories abound on the location of code switching; i.e., whether intrasentential code switching is alternational or insertional (the insertion of a secondary lexical item into a primary base structure). Determination of which language is primary by an automatic system is sometimes based on the consideration of left-to-right parsing where the language of the first words encountered is chosen [13].

A speech recognizer for mixed language has the same challenges and designs as a multilingual recognizer, with the variation that words in both languages appear in the same sentence and are spoken by the same speaker. While multilingual and multi-accent acoustic modeling have been discussed above, this section focuses on discussing the second stage of understanding mixed-language sentences.

To understand mixed-language sentences, a system needs to be able to establish the equivalence of the inserted secondary words in the primary base language and to predict the location of code switching. To achieve the first objective,

mixed-language understanding uses translation disambiguation methods to convert these queries into monolingual queries [3]. A secondary language word is first glossed and then translated into the correct primary language word by looking at the context in the sentence. The main technical challenge is ranking the translation candidates. In [7], the authors investigated various strategies of using context words in a sentence to disambiguate translation candidates.

The challenge of mixed-language processing remains keen as the location of code switching is often considered unpredictable. Fung and Cheung [7] suggest predicting the code-switch location either from empirical data or from linguistic rules. Researchers may take the tree structure in a mixed language into account to better predict code switching. One such approach as described in [3] is to assume that secondary-language words are only inserted as nouns or noun phrases and thus train a mixed-language model.

A special case of mixed-language understanding is the understanding of mixed-language named entities. Names for locations, places, organizations, and people are a mixture of phonetic transliterations and semantic translations, such as "Victoria Falls →維多利亞瀑布 (wei duo li ya pu bu)." Person names are often transliterated. Location names are partly transliterations and partly translations. Organization names are more likely to be translated. Statistical methods, based on frequencies of the named entities, are employed to solve these problems [19].

BRIDGING THE GAP BETWEEN LANGUAGE AND TECHNOLOGY EXPERTISE

Speech processing systems have been developed for a number of domains and languages in recent years. However, in spite of well-developed toolkits, it is still a skilled task requiring significant effort from trained individuals. Therefore, one of the most important trends is the need to support multiple languages on demand or, in the best case, almost instantly. New algorithms and strategies are required that support the rapid adaptation of speech processing systems to new domains and languages.

The SPICE project [29] aims to overcome this limitation by providing innovative methods and interactive Web-based tools to enable novice users to develop speech processing systems, to collect appropriate data for building these models, and to evaluate the results allowing for iterative improvements. It leverages the GlobalPhone [27] and FestVox (Black and Lenzo, http://festvox.org) projects and implements bootstrapping techniques that are based on extensive knowledge and data sharing across languages as well as sharing across system components. An interface to the Web-based SPICE tools has been designed to accommodate all potential users. It allows for speech and text data harvesting and archiving, soliciting basic language information, logging activity, user profiles, development projects, and the automatic construction of speech recognition and speech synthesis including acoustic model bootstrap from multilingual models, vocabulary selection, language model generation, and pronunciation generation.

User studies were carried out to indicate how well speech systems can be built, how well the tools support development efforts, and what must be improved to create even better systems. While the SPICE tools have been used previously to bootstrap speech recognition systems in Afrikaans, Bulgarian, and Vietnamese within a 60-day timeframe, they recently targeted the parallel development of speech processing components for a broader range of languages within a six-week hands-on lab course at Carnegie Mellon University [29]. Students were required to rely solely on the SPICE tools and report back on problems and limitations of the SPICE system. Results indicate that it is feasible to build various end-to-end speech translation systems including speech recognition, speech synthesis, and a statistical translation system in new languages for small domains within the framework of a six-week course.

SPICE will hopefully revolutionize the system development process in the future. Archiving the data from many cooperative native users will significantly increase the repository of languages and resources. Data and components for new languages will become available at large to let everyone participate in the information revolution, improve the mutual understanding, bridge language barriers, and thus foster educational and cultural exchange.

DISCUSSION AND OPEN ISSUES

In this article we have provided a systematic survey of state-of-the-art multilingual technologies and applications. We have discussed the challenges for multilingual systems; namely, the lack of language resources and expertise, the need for rapid development in new languages, and consequently the demand for language-independent approaches that can be applied to multilingual applications even with only little resources and without any stringent need for language-specific knowledge. We have surveyed the latest trends in multilingual acoustic and language modeling, dialog systems, speech summarization, and speech translation. We have shown that the most popular approach in these areas is statistical modeling with machine-learning methods, separating language-specific resources from core language-independent engines. However, language-specific resources such as parallel corpora; named entity lists, parsers and taggers, phrase tables, and pronunciation dictionaries continue to be essential components for the systems' performance. Therefore, all multilingual applications will benefit from finding efficient methods to acquire these language-specific resources in an automatic or semi-automatic way. We also foresee other research directions for multilingual applications. For example, multilingual emotional speech identification and classification is only at a nascent stage, where large databases of emotional speech are being collected in different languages. Spoken summarization is currently restricted to one or two languages, and it is feasible to extend this to multilingual systems. Large-scale multilingual systems with simultaneous processing of multiple languages, although rare today, are likely to be of higher demand in today's world of

Internet search and globalization. Commercial organizations such as Google and Microsoft are unique in their possession of the world's largest database of multilingual input and output data. The collaboration between commercial, government, and academic communities will ultimately revolutionize the field of multilingual spoken language processing.

AUTHORS

Pascale Fung (pascale@ece.ust.hk) received her master's and Ph.D. in computer science from Columbia University in 1993 and 1997, respectively. She also studied at Ecole Centrale Paris and Kyoto University and was formerly a researcher at AT&T Bell Labs and LIMSI/CNRS in France. She joined the Hong Kong University of Science and Technology (HKUST) in 1997 and is currently an associate professor in electronic and computer engineering. She cofounded the Human Language Technology Center and is the director of InterACT at HKUST. Her research interests include multilinguality in both speech and language processing. She is a Senior Member of the IEEE and a board member of SIGDAT, the Association of Computational Linguistics.

Tanja Schultz (tanja@cs.cmu.edu) received her master's and Ph.D. in computer science from Karlsruhe University, Germany, in 1995 and 2000, respectively, and a master's in mathematics and sports from the University of Heidelberg, Germany, in 1990. She joined Carnegie Mellon University in 2000 and is an assistant research professor at the Language Technologies Institute and serves as associate director of InterACT. Since 2007 she also has been a full professor at the Computer Science Department of Karlsruhe University. Her research focuses on multilingual speech processing with a special emphasis on rapid language adaptation.

REFERENCES

[1] A. Black and T. Schultz, "*Speaker clustering for multilingual synthesis*," presented at MULTILING ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing, Stellenbosch, South Africa, Apr. 2006.

[2] P. Cheung and P. Fung, "Unsupervised learning of a spontaneous and colloquial speech lexicon in Chinese," *Int. J. Speech Technol.*, vol. 7, no. 2, pp. 173–178, Apr. 2004.

[3] P. Cheung and P. Fung, "Translation disambiguation in mixed language queries," *Mach. Translat. J.*, vol. 18, no. 4, pp. 273–251, Dec. 2005.

[4] M. Davel and E. Barnard, "The efficient generation of pronunciation dictionaries: Human factors during bootstrapping," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Jeju, Korea, 2004, pp. 2797–2800.

[5] C. Fügen, S. Stüker, H. Soltau, F. Metze, and T. Schultz, "Efficient handling of multilingual language models," in *Proc. Workshop Automatic Speech Recognition Understanding (ASRU-2003)*, St. Thomas, Virgin Islands, Dec. 2003, pp. 441–446.

[6] P. Fung, "*Multilingual language processing*," presented at IEEE/ACL Workshop on Spoken Language Technology, Palm Beach, Aruba, 2006.

[7] P. Fung and P. Cheung, "Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpora," in *Proc. 20th Int. Conf. Computational Linguistics (COLING 2004)*, Geneva, Switzerland, Aug. 2004, article 1051.

[8] P. Fung and G. Ngai, "One story one flow: Hidden Markov story models for multilingual and multi-document summarization," *ACM Trans. Speech Lang. Processing*, vol. 3, no. 2, pp. 1–16, July 2006.

[9] D. Gildea, "Loosely tree-based alignment for machine translation," in *Proc. 41st Annu. Meeting Association for Computational Linguistics*, Sapporo, Japan, July 2003, pp. 80–87.

[10] M. Honal and T. Schultz, "Automatic disfluency removal on recognized spontaneous speech—Rapid adaptation to speaker-dependent disfluencies," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Philadelphia, PA, Mar. 2005, pp. 969–972.

[11] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "A statistical approach to automatic speech summarization," *EURASIP J. Applied Signal Processing*, vol. 2003, no. 2, pp. 128–139, 2003.

[12] C. Huang, E. Chang, J. Zhou, and K.-F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition," in *Proc. ICSLP 2000*, Beijing, China, Oct. 2000, pp. 818–821.

[13] A.K. Joshi, "*Processing of sentences with intrasentential code-switching*," in: Nat. Language Parsing, D.R. Dowty, L. Karttunen, and A.M. Zwicky, Eds. Cambridge, UK: Cambridge Univ. Press, pp. 190–205, 1985.

[14] S. Kanthak and H. Ney, "Multilingual acoustic modeling using graphemes," in *Proc. European Conf. Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 1145–1148.

[15] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of designed algorithm based upon the generalized probabilistic descent method," *Proc. IEEE*, vol. 86, no. 11, pp. 2345–2373, Nov. 1998.

[16] S. Khudanpur and W. Kim, "Contemporaneous text as side-information in statistical language modeling," *Comput. Speech Lang.*, vol. 18, no. 2, pp. 143–162, 2004.

[17] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in: *Proc. European Conf. Speech Communication and Technology*, Geneva, Switzerland, Sept. 2003, pp. 3141–3144.

[18] K. Kirchhoff and D. Vergyri, "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition," *Speech Commun.*, vol. 46, no. 1, pp. 37–51, May 2005.

[19] K. Knight and J. Graehl, "Machine transliteration," *Computat. Linguistics*, vol. 24, no. 4, pp. 599–612, 1998.

[20] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Philadelphia, PA, Mar. 2005, pp. 1–4.

[21] E. Levin and R. Pieraccini, "A stochastic model of computer-human interaction for learning dialogue strategies," in *Proc. Eurospeech'97*, Rhodes, Greece, 1997, pp. 1883–1886.

[22] Y. Liu, and P. Fung, "State-dependent tied mixtures with pronunciation modeling for spontaneous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 4, pp. 351–364, July 2004.

[23] Y. Liu and P. Fung, "Multi-accent Chinese speech recognition," in *Proc. Interspeech 2006*, Pittsburgh, PA, Sept. 2006, pp. 133–136.

[24] C.Y. Ma and P. Fung, "Using English phoneme models for Chinese speech recognition," in *Proc. Int. Symp. Chinese Spoken Language Processing*, Hong Kong, pp. 80–82, Dec. 1998.

[25] S. Maskey and J. Hirschberg, "Summarizing speech without text using hidden Markov models," in *Proc. NAACL*, 2006, pp. 89–92.

[26] T. Schultz and K. Kirchhoff, Eds., *Multilingual Speech Processing*. New York: Academic Press, Apr. 2006.

[27] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1–2, pp. 31–51, Aug. 2001.

[28] T. Schultz, A. Black, S. Vogel, and M. Woszczyna, "Flexible speech-to-speech translation systems," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 2, Mar. 2006.

[29] T. Schultz, A. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based tools for rapid language adaptation in speech processing systems," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2125–2128.

[30] Y.-C. Tam, I. Lane, I., and T. Schultz, "Bilingual-LSA based LM adaptation for spoken language translation, " in *Proc. ACL*, Prague, Czech Republic, June 2007, pp. 520–527.

[31] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kiramura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, Istanbul, Turkey, 2000, pp. 1315–1318.

[32] U. Übler, M. Schüßler, and H. Niemann, "Bilingual and dialectal adaptation and retraining," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, Nov. 1998, pp. 1815–1818.

[33] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," in *Computat. Linguistics*, vol. 23, no. 3, pp. 377–404, Sept. 1997.

[34] K. Yamada and K. Knight, "A syntax-based statistical translation model," in *Proc. 39th Annu. Meeting Association for Computational Linguistics*, Toulouse, France, July 2001, pp. 523–530.

[35] J. Zhang and P. Fung, "Speech summarization without lexical features for Mandarin broadcast news," in *Proc. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, Apr. 2007, pp. 213–216.

[36] J. Schroeder and P. Koehn, "The University of Edinburgh system description," in *Proc. IWSLT 2007, International Workshop on Spoken Language Translation*, Trento, Italy, Oct. 2007. SP