TOWARDS LANGUAGE PORTABILITY IN STATISTICAL SPEECH TRANSLATION

Alex Waibel, Tanja Schultz, Stephan Vogel,
Christian Fügen, Matthias Honal, Muntsin Kolss, Jürgen Reichert, Sebastian Stüker

Interactive Systems Laboratories,
Carnegie Mellon University and Karlsruhe University

1. INTRODUCTION

Speech translation has made significant advances over the last years with several high-visibility projects focussing on diverse languages in restricted domains (e.g. C-Star, Nespole, Babylon). When addressing spontaneous conversational speech translation the solution cannot be expected to be a mere connection of ASR and MT due to the peculiarities of spoken language, and the disfluent, fragmentary nature of spontaneous speech. Furthermore, while speech recognition emerged to be rapidly adaptable to new languages in large domains, translation still suffer from the need of hand-crafted grammars for interlingua-based approaches or the lack of large parallel corpora for statistical machine translation. Both facts prevent the efficient portability of speech translation systems to new languages and domains. We believe that we can overcome today's limits of language and domain portable conversational speech translation systems by relying more radically on learning approaches and by the use of multiple layers of reduction and transformation to extract the desired content in another language. Therefore, we cascade several stochastic source-channel models as shown in figure 1 that extract an underlying message from a corrupt observed output. The three models effectively translate: (1) speech to word lattices (ASR), (2) ill-formed fragments of word strings into a compact well-formed sentence (Clean), and (3) sentences in one language to sentences in another (MT). In this paper we present results of our research efforts towards rapid language portability of all these components.

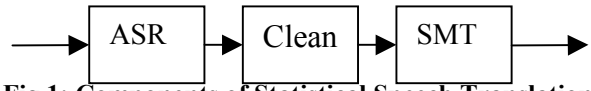

Fig.1: Components of Statistical Speech Translation

2. MULTILINGUAL ASR

Over the last couple of years we accumulated considerable experience in language adaptation techniques in acoustic modeling [Schultz]. Based on our multilingual data collection efforts, such as GlobalPhone, we combine the data of various languages to train a global, language independent phone set and adapt the corresponding acoustic models to new target languages. Our rapid adaptation techniques enables us to bootstrap acoustic models in a new language with very limited training data. Recently, we extended our efforts in rapid bootstrapping of LVCSR recognition system to the automatic generation of pronunciation dictionaries in a multilingual setting. We applied grapheme-based rather than phoneme-based models and thus derive pronunciations directly from the given transcription. Our results for English, German, and Spanish show that for languages with reasonable letter-to-sound relation this approach gives comparable results [Killer].

3. DISFLUENCY CLEANING

Spontaneous spoken speech usually contains disfluencies such as filler words, repairs or restarts which do not contribute to the meaning of the spoken utterance. This cause sentences to be ill-formed, longer, and thus harder to process for translation. We developed a cleaning component that automatically removes those parts from the speech recognition output which do not belong to the utterance originally intended by the speaker. Our approach is based on a noisy-channel model and its development requires no linguistic knowledge, but only annotated texts. Therefore, it has large potential for rapid deployment and the adaptation to new target languages.

We adopt Shriberg's definition [Shriberg] in which a disfluency consists of the reparandum (the words which will be repeated or corrected), the interruption point, the interregnum (silence or editing term indicating the existence of a reparandum), and the repair. Following this definition, disfluencies can be corrected by the deletions of the interregnum and the reparandum.

3.1 Noisy-channel approach for disfluency cleaning

The cleaning component is based on a noisy-channel approach, a basic concept of SMT [Wang] that we adapted to the problem of disfluency cleaning, by assuming that "clean" i.e. fluent speech gets passed through a noisy channel. The channel adds noise to the clean speech and creates "noisy", i.e. disfluent speech as output. Given a "noisy" string N the goal is to recover the "clean" string $\hat{C}$ such that $p(\hat{C}|N)$ becomes maximal. Using Bayes Rule, this problem can be expressed as:

$$\hat{C} = \arg\max_{C} p(C \mid N) = \arg\max_{C} p(N \mid C)\, p(C)$$

where the probability $p(C)$ denotes the language model probability for the fluent string; $p(N|C)$ is the probability

that the noisy channel generates N as output given C as input. In terms of SMT the latter probability is referred to as the translation model. Like in SMT we use alignments to establish correspondences between the positions of the source and the target sentences, however in the case of disfluency cleaning only deletions of words need to be considered. We then search for the most likely target language sentence given a sentence in a source language. This search takes all possible hypotheses into account which can be generated from the source sentence by deletion. In order to assign probabilities to these hypotheses, a number of models for different properties of disfluencies are used as described below.

Assuming that each target sentence is generated from left to right, the alignment a_j defines whether the word n_j in the source sentence is deleted or appended to the target sentence. Let J be the length and n_j the words of the source sentence N, I the length and c_j the words of the target sentence C and m the number of deletions (of contiguous word sequences) which are made during generation of the target sentence. Then we can introduce an alignment a_j for each word n_j and rewrite $P(N|C)$ as:

$$p(N \mid C) = p(J \mid I)\, p(m \mid J, C) \sum_{a_1^J} p(n_1^J, a_1^J \mid c_1^J, I, J, m)$$

The probability $p(n_1^J, a_1^J \mid c_1^J, I, J, m)$ can be decomposed into a product of probabilities over all source words n_j. In our system we use five different models which contribute to these probabilities and are combined by a weighted sum. Each model assigns a translation probability to a word: (M1) The length of the deletion region of a disfluency, (M2) the position of a disfluency, (M3) the length of the deletion region of a disfluency with a fragment at the end of the reparandum, (M4) the context of a potentially disfluent word, (M5) the information about the deletions of the last two words preceding a potentially disfluent word. The models (M1), (M2) and (M3) reflect important properties for disfluency identification as outlined in [Honal]. Models (M4) and (M5) take into account that the local context is often helpful to determine the deletion region of a disfluency.

3.2 Experimental Results

The probability distributions for the models encoding the features enumerated above are obtained from the training data using relative frequencies. All experiments are conducted on spontaneously spoken dialogs in English and, in order to demonstrate the feasibility of rapid adaptation, additionally on the spontaneous Mandarin Chinese CallHome corpus. The English corpus was split into 10 disjoint test sets of 10% corpus size, the corresponding 90% remainder of the corpus were used for training. The presented results on EVM are averaged over the 10 test sets. For the MCC corpus we used the predefined splitting into a training, development, and evaluation set. The results on the MCC are reported on the evaluation test set.

Model	*Hits*	*False Positive*
M1	-7.9	-5.9
M2	+10.1	+21.2
M3	+2.6	+8.8
M4	+174.1	-90.3
M5	+15.9	+264.6

Table 1: Contribution of all five models to the baseline

The effect of all the five models is summarized in table 1. The most remarkable effect on the overall performance gain results from model (M4) which considers the context of a potentially disfluent word. This can be easily explained for filler words, since it allows to discriminate between the deletion of the word "well" in the context "Well done!" and "Alright, well, this is a good idea". The impact of (M1), (M2), and (M3) is a slight increase of the number of hits at the cost of a slight increase or decrease of the number of false positives. Model (M5) causes a huge number of false positives and was therefore disregarded in the best system.

	Hits	*False Positive*	*Recall*	*Precision*
English	853.3(1102.7)	92.4	77.2%	90.2%
Chinese	1486(3008)	448	49.4%	78.8%

Table 2: Disfluency cleaning results for two languages

Overall a recall of 77.2% and a precision of 90.2% was obtained for English dialogs as shown in table 2. Almost no effort was required for the adaptation to Mandarin Chinese. The same algorithms and the same statistical models were used, only the weighting parameters for the models were adjusted. We achieved 49.4% recall and 76.8% precision on the Mandarin corpus. In conclusion, our approach has several advantages for the development of a cleaning system: (1) Language portability: no linguistic knowledge is needed, but only text containing annotated disfluencies. (2) Granularity: rather than rules, statistical models are used to make decisions about deletions which allow for case-to-case decisions depending on a number of features. (3) Flexibility: easy integration of new models that make use of disfluency properties yet to be investigated.

4. STATISTICAL MT USING ENGLISH AS INTERLINGUA

This section describes the Error Driven Translation Rule Learning (EDTRL) system that uses a form of augmented, formalized English as an interlingua to translate from a source language into a target language. EDTRL eliminates the drawbacks of interlingua- and data-driven approaches since (1) it avoids the need for an explicit, handcrafted interlingua specification, and (2) tackles the "Parallel Data

Sparseness Problem" which limit the pure data-driven systems. As a result this approach is well suited for the rapid adaptation of translation from and to new languages.

4.1 Formalized English as Interlingua

In the last couple of years, the translation from and to English made significant progress and nowadays various translation tools together with large parallel corpora are available. However, the situation drastically changes if one looks for support to translate from and to languages other than English. The intuitive solution to this problem is to cascade two translators using English as the intermediate language. The main problem of this solution is the multiplication of translation errors. Therefore, the focus of the EDTRL approach is to reduce this multiplication effect. The basic design idea is to preserve translation alternatives and incorporate additional knowledge about the structure and content of a sentence.

With respect to preserving translation alternatives, we examined three methods, (1) keep n-best list of complete translations, (2) keep n-best word or phrase alternatives, and (3) keep the full lattice. In the first method up to *n* alternative translation hypotheses are produced and passed to the second translation step. To guarantee fast decoding, *n* needs to be kept small. This approach did not improve the translation in our experiments. In the second method the single best hypothesis from the first translation step was selected, but augmented by adding alternative words or phrases, which have high translation probabilities. This strategy results in a noticeable improvement in the translation performance. In the third method all alternatives, i.e. the full translation lattices is passed on to the second step which increases the search space considerably. This method performed best in our experiments.

Besides preserving translation alternatives we incorporated five additional knowledge sources concerning the structure and semantic of a sentence:

1. Morphological Analyzer: Analyze the English word form based on the WordNet ontology [WordNet] and determine its base form and derivation rule.
2. Sense Guesser: Determines the meaning of a word in a given context based on the sense hierarchy from WordNet.
3. Synonym Generator: Provides a lists of synonyms gathered from WordNet.
4. Part-of-Speech Tagger: provides POS-tags defined by the tag set described in [Brill] and trained on the tagged Brown Corpus.
5. Named Entity Tagger: detects named entities.

Together with the translation alternatives, these knowledge sources form the interlingua for the EDTRL system. In order to cut out less relevant information or convert to more common phrases, the intermediate English was additionally formalized by some rules, e.g. "Please give me X" is transformed to "Give me X, please".

4.2 Training and Translation Process

EDTRL is based on statistical transfer rules, learnt from automatically tagged bilingual corpora. Annotation is only required for English, word alignment models are used to project this information into the other language. The knowledge sources operate on English only and are independent from the input and output language. The use of probabilistic translation rules allow to add new rules, model exceptions, track and correct translation errors.

Statistical Alignment

In a first step a word alignment (IBM1 or modified IBM2) is performed, which builds the basis for the phrase alignment in the second step. The phrase alignment simultaneously joins similar regions on the word alignment matrix and splits the matrix in smaller parts. The split and join operations use normalized probabilities from the word alignment and the language models. The result of both steps is a collection of partitions of the word alignment matrix and their probabilities.

Rule Generation and Selection

Based on the alignment, the optional dictionaries, as well as the semantic and morphologic knowledge, translation rules are generated. Rules are of the form:

Cond1 | *Cond2* | ... → *Templ1* | *Templ2* | ...

where *Cond* can be a word or phrase containing attribute classes and *Templ* is a template which has to be instantiated during the translation process. Probabilities are assigned to both, *Cond* and *Templ*. Most attribute classes are part of a hierarchy which allows enforcing a match by traversing the tree up to a more common representation while at the same time decreasing the rule score. A set of meta-rules describes the construction process. In order not to restrict the rule set, the efficiency of each rule is determined on a validation set.

The Translation Process

The translation process tries to match and instantiate rules along the input utterance. This results in a search tree which needs to be pruned to limit the size.

4.3 Experiments

To evaluate the concept of English as an interlingua we chose Chinese as input and Spanish as output language, since, in spite of the widespread use of these languages, comparatively few direct Chinese-Spanish translations are available. We trained the EDTRL system for Chinese to English (C→E), English to Spanish (E→S) and Chinese to Spanish (C→S). The output of the C→E system was then used as input for the E→S system. A second cascaded translation was performed, but this time using the formalized English as intermediate step. Additionally, we trained a statistical MT system [Vogel 2003] on the same language pairs and also cascaded the C→E and E→S

translations to generate a C→E→S translation in comparison to a direct C→S translation. To evaluate the translation quality we used NIST MTeval v9c [MTeval]. The results are listed in Table 4. For comparison, we give also the results for Systran's publicly available translation system [Systran].

The SMT system and the EDTRL system both use the same bilingual training corpus, while the EDTRL system uses additional dictionaries for initialisation. An additional difference is the handling of punctuation. While EDTRL ignores punctuation marks, SMT treats them as normal words. In the reported experiments the EDTRL system does not make use of the Sense Guesser, the Named Entity Tagger, and full lattice.

Train (Test)	*Chinese*	*English*	*Spanish*
sentences	162316 (506)	162316 (506)	6027
-unique	96074 (497)	97500 (503)	5934
-avg. length	7.0 (7.3)	7.5 (7.5)	9.8
words	1134417 (3681)	1216207 (3779)	58834
vocabulary	13793 (954)	16224 (843)	4651
-singletons	4745 (590)	6705 (523)	2370
-unseen	(29)	(22)	

Table 3: Training (Test in parentheses) corpora

The data for these experiments were taken from the Basic Travel Expression Corpus (BTEC), a multilingual collection of conversational phrases in the travel domain [Takezawa]. Table 3 shows the training and test material for Chinese, English and Spanish phrases. Since only a subset of 6027 phrases was available for Spanish, only the corresponding parallel phrases were used to train the E→S and C→S systems. The scores were calculated using 16 English and in average 3-4 Spanish reference translations.

NIST-Score	*EDTRL*	*SMT*	*Systran*
C → E	6,73	7,35	5,74
E → S	5,17	4,57	6,06
C → S	2,84	3,04	-
C → E → S	3,34	2,60	2,84
C → E_{IL} → S	3,52	-	-

Table 4: Translation Results

The higher scores of the statistical systems on Chinese to English compared to translations to Spanish result from the facts, that much more training material was used and the evaluation was performed with a higher number of references. For Systran both numbers are closer, while it seems the E→S system is slightly better. Surprisingly, the cascaded EDTRL systems outperform the directly trained system. Using augmented and formalized English as an interlingua in the EDTRL system is shown to yield improvements over the cascaded approach.

5. CONCLUSION

In this paper we presented an approach towards the tighter coupling of statically based speech translation that uses multiple layers of reduction and transformation by cascading several stochastic source-channel models. This approach more radically relies on learning techniques to overcome today's limits of language and domain portable conversational speech translation systems. The disfluency cleaner for English achieved a recall of 77.2% and a precision of 90.2%. The same algorithms and models were effortless adapted to Mandarin Chinese giving 49.4% recall and 76.8% precision. The results on translation suggest that MT systems can be successfully constructed for any language pair by cascading multiple MT systems via English. Moreover, end-to-end performance can be improved, if the interlingua language is enriched with additional linguistic information that can be derived automatically and monolingually in a data-driven fashion.

6. REFERENCES

[Brill] E. Brill, A Case Study in Part of Speech Tagging, *Association for Computational Linguistics, 1995*.

[Honal] M. Honal and T. Schultz, Correction of Disfluencies in Spontaneous Speech using a Noisy-Channel Approach. Proc. Eurospeech 2003, Geneva, Switzerland.

[Killer] M. Killer, S. Stüker, T. Schultz, Grapheme-based Speech Recognition, Proc. Eurospeech 2003, Geneva, Switzerland.

[MTeval] http://www.nist.gov/speech/tests/mt/ NIST MT evaluation kit version 9.

[Schultz] T. Schultz and A. Waibel, Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition, Speech Communication, Vol. 35, August 2001.

[Shriberg] E. Shriberg, Preliminaries to a Theory of Speech Disfluencies, PhD-Thesis, University of California at Berkley, 1994.

[Systran] http://www.systranbox.com/systran/box

[Takezawa] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. *Proc. of LREC 2002*, pp. 147–152, Las Palmas, Canary Islands, Spain, May 2002.

[Vogel] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel. "The CMU Statistical Translation System." Proc. of the MT Summit IX. New Orleans, LA. September 2003.

[Wang] Y. Wang, A. Waibel, Decoding Algorithm in Statistical Machine Translation, Proc. of the 35th Annual Meeting of the ACL, 1997

[WordNet] http://www.cogsci.princeton.edu/~wn/