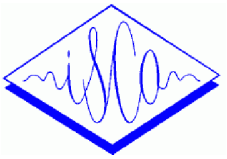
ISCA Archive
http://www.isca-speech.org/archive



International Workshop on Spoken Language Translation (IWSLT) 2007
Trento, Italy
October 15-16, 2007

The University of Maryland Translation System for IWSLT 2007

Christopher J. Dyer

Department of Linguistics
University of Maryland
redpony@umd.edu

Abstract

This paper describes the University of Maryland statistical machine translation system used in the IWSLT 2007 evaluation. Our focus was threefold: using hierarchical phrase-based models in spoken language translation, the incorporation of sub-lexical information in model estimation via morphological analysis (Arabic) and word and character segmentation (Chinese), and the use of n-gram sequence models for source-side punctuation prediction. Our efforts yield significant improvements in Chinese-English and Arabic-English translation tasks for both spoken language and human transcription conditions.

1. Introduction

In recent years, phrase-based techniques have become the predominant paradigm in statistical machine translation. In addition to traditional phrase-based models [1], models that incorporate hierarchical structure [2] and syntactic relationships [3] also rely heavily on phrases of contiguous words to function as anchors in the translation process. The spoken language translation task explored in the IWSLT workshop poses additional challenges not found in text based translation using phrase-based systems. We explore three of these challenges and offer strategies for coping with the problems they present:

1. **Imperfect recognition**. Ney [4] demonstrated that when translating ASR output, utilization of information beyond what is contained in the single best transcription hypothesis can lead to better translation quality. More recently, Bertoldi et al. [5] showed that by modeling the transcription hypothesis space with confusion networks, highly efficient translation using a conventional phrase-based translation model is possible. We present results showing improvements by extending this technique to *hierarchical* phrase-based models.

2. **Data sparseness**. Although increasing amounts of parallel data are becoming available, parallel corpora of spoken language and/or spoken language transcripts remain rare. We address the problems caused by data sparseness by incorporating knowledge about the internal structure of words in Chinese and Arabic to improve estimation of translation models.

3. **Missing source-side punctuation**. ASR output generally does not contain any sort of punctuation; however, correct punctuation is necessary in target language translations. Furthermore, source side punctuation can play an important role in translation of other words in a sentence. We show that by predicting punctuation on the source side, translation quality improves.

In the following sections we review hierarchical phrase-based translation and describe how an ASR word lattice approximated by a confusion network can be translated efficiently. We then present results showing the benefits of using sub-lexical information when constructing translation models as well as the benefits of carrying source-side punctuation prediction and translating the resulting representation. We conclude with a summary and ideas for future research.

2. Hierarchical phrase-based translation

Recently, hierarchical phrase-based translation models (HPBTMs) were introduced to address some of the perceived shortcomings of standard phrase-based models [2]. A hierarchical model generalizes phrase-based translation models by allowing phrase pairs to contain variables. Like phrase pairs in the classical models, the synchronous grammar rules in HPBTMs are learned automatically from aligned, but otherwise unannotated, training corpora. For details about the rule extraction algorithm, refer to [2]. Formally, HPBTMs form a synchronous context-free grammar, and the translation process is equivalent to parsing the source sentence with one side of the grammar, thereby inducing a tree in the target language.

The induced grammar rules consist of pairs of strings of terminals and non-terminals in the source and target languages, as well one-to-one correspondences between non-terminals on the source and target side of each pair (shown as indexes in the examples below). Thus they encapsulate not only meaning translation (of possibly discontinuous spans), but also typical reordering patterns. For example, the following two rules were extracted from the Arabic → English por-

tion of the BTEC training data provided by IWSLT 2007:[1]

$$\mathrm{X} \rightarrow \langle t{*}hb\ \mathrm{X}_{\boxed{1}}\ AlY\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ goes\ to\ \mathrm{X}_{\boxed{2}} \rangle \quad (1)$$

$$\mathrm{X} \rightarrow \langle Al\ \mathrm{X}_{\boxed{1}}\ Al\ HmrA', the\ red\ \mathrm{X}_{\boxed{1}} \rangle \quad (2)$$

Rule (1) expresses that the finite Arabic verb *t∗hb* (English, *goes*) generally proceeds its subject, and the destination is given by a prepositional phrase begun with *AlY* (English, *to*). Rule (2) shows that the definite marker *Al* and adjective *HmrA'* follow a modified definite noun whereas the corresponding English adjective *red* precedes what it modifies and there is only a single definite article. Although the rules given here correspond to syntactic constituents, this is accidental. The induced grammars make use of only a single non-terminal category and variables may or may not correspond to linguistically meaningful objects.

Given a synchronous grammar G, the translation process is equivalent to parsing an input sentence with the source side of G and thereby inducing a target sentence. The translation model features are combined in a log-linear framework and efficient dynamic programming parsers exist based on the CKY+ algorithm, which permits the parsing of rules that are not in Chomsky normal form [6]. We have further extended this algorithm to admit input that is in the form of a confusion network, as described in [7]. To incorporate target language model probabilities into the model, which is crucially important for translation quality, the grammar is intersected during decoding with an n-gram language model. This intersection significantly increases the effective size of the grammar, and so a beam-search heuristic called *cube pruning* is used, which has been experimentally determined to be nearly as effective as an exhaustive search but far more efficient [2].[2]

2.1. Confusion networks

A confusion network (CN) is a directed acyclic graph with a single start node and single end node and which has the property that every path passes through every node exactly once [8].[3] Although originally developed to maximize the posterior probability of a word in ASR systems that seek to maximize the *sentence* posterior probability, their compact representation has made them attractive for a variety of other applications. Efficient algorithms exist for using them as input for translation in a traditional phrase-based decoder [5] as well as a hierarchical phrase-based decoder [7].

[1] Arabic examples in this paper are given in the Buckwalter transliteration. http://www.qamus.org/transliteration.htm

[2] The full list of features used in all experiments is: $P(e|f)$, $P(f|e)$, $P_w(e|f)$, $P_w(f|e)$, $P_{LM}(e)$, $P_{CN}(f)$ (probability of the source phrase in the confusion network), and the count features: $C(e)$, and counts of the number of rules with one and two non-terminals respectively. A trigram language model trained on the target side of the training corpus with modified Kneser Ney smoothing was used for all experiments.

[3] Confusion networks are also commonly referred to as pinched lattices, meshes, and sausages.

	1-best	Full-CN
BLEU	17.25	17.72
TER	65.95	63.51

Table 1: Incorporation of alternate transcription hypotheses into the translation of Chinese.

The development and test word lattices provided for the IWSLT 2007 evaluation were converted into confusion networks using SRI's `lattice-tool` with no pruning.[4] The resulting confusion networks contain at least as many paths as were contained in the original word lattice.

2.2. Experimental results

Table 1 shows the effect of translating a 1-best hypothesis ASR hypothesis compared with translating a full CN. The system was tuned on the DEV4 development set, repunctuated as described in Section 4. On this evaluation set, there is a marginally significant improvement in BLEU score for using the full confusion network, and a significant reduction in TER ($p < .05$).[5]

3. Sparse training data: Smoothing word models with sublexical information

Phrase-based machine translation models, with few exceptions, treat linguistic elements (e.g. words, phrases) that are not orthographically identical as statistically independent ([1], [2], inter alia; for a counterexample see [9]). This naive assumption is made for the sake of computational tractability and because the precise nature of the relationship between word forms is highly language dependent (for example, in English the plural is commonly marked by adding the *-s* suffix to a base noun, whereas in Arabic it is commonly formed by realizing the consonantal root of the base noun with a different vocalic template). However, the orthographic independence assumption comes at a cost: when building a translation model involving languages with large numbers of distinct orthographic tokens (such as is the case with morphologically complex languages), there are more parameters to estimate and therefore larger quantities of training data are necessary to achieve reasonable performance.

An often-used strategy for mitigating the negative effects of this independence assumption involves clustering orthographic forms into equivalence classes and building translation models over the clusters rather than the more numerous raw word forms. Although conceptually quite significant, the commonly utilized forms of this strategy are so widespread they receive virtually no mention. Two examples of this are case normalization (lowercasing or other related strategies in

[4] Refer to http://www.speech.sri.com/projects/srilm/ for more information.

[5] All statistical significance tests reported in this paper were carried out using the bootstrap sampling method described in [14].

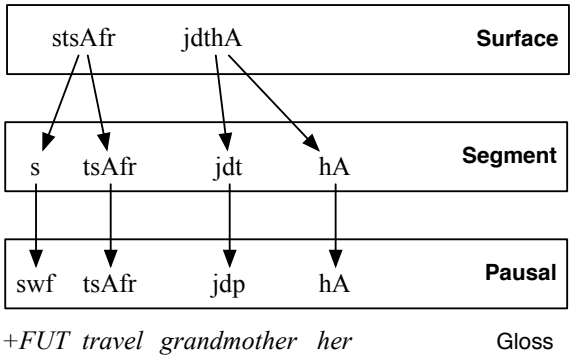


Figure 1: Two-phase Arabic morphological analysis and normalization of the fragment *stsAfr jdthA*, English *her grandmother will travel*.

languages whose alphabets distinguish upper and lower case forms) and word segmentation (splitting punctuation away from words so that tokens next to a punctuation character are identical to the freestanding forms, or splitting sentences written without whitespace into constituent words).

We use high-level linguistic knowledge about the structure of Arabic and Chinese word forms to motivate strategies for building translation models with fewer parameters that break down the orthographic independence assumption. This in turn yields improvements in translation quality, which we approximately quantify with BLEU score and statistics about the number of untranslated words in the output.

3.1. Arabic segmentation and normalization

In Arabic, a large class of function words (including conjunctions, prepositions, possessive and object pronouns, and tense morphemes) are attached to the words they follow or proceed as clitics. Since any word may have several clitics attached, and since these clitics generally can be translated independently of their host stem, splitting them into independent elements is a reasonable approach to reducing sparse data. Although splitting Arabic words into their constituent morphemes has been done in the past (see for example, [10] and especially the "English-like" condition described in [11]), the approach taken here goes a step further and attempts to normalize the recovered segments into the form they would have if they were stand-alone morphemes. This is schematized in Figure 1. The following strategies are used for splitting and normalizing Arabic tokens:

- Short vowels and other diacritics are removed.
- The use of *ya'* and *alif maqsura* is normalized.
- *Hamza* and *madda* are stripped from *alif* since their use is inconsistent.
- Tokens are analyzed using the Buckwalter Morphological Analyzer (BAMA). Clitics are split from their host stems if the segmentation hypotheses delivered by BAMA agree with each other (the identity of the stems and affixes do not have to agree).
- Suffixes indicating number or gender features are left attached.
- Spelling changes resulting from cliticization are undone (e.g., restoring the pausal feminine singular ending, $t \rightarrow p$).
- The future tense prefix *s-* is normalized to the freestanding particle *swf*.

Arabic morphological processing was carried out on a token-by-token basis, which meant that even it could be applied to word lattices where the context of a given token is a network of possibilities rather than a simple string. Refer to Figure 2 for examples of the normalization and segmentation used.

Compound	Meaning	Character glosses	
便携式	portable	便	convenient
		携	carry
		式	kind
名人	celebrity	名	famous
		人	person
无声电影	silent film	无	no
		声	voice
		电	electric
		影	film

Figure 3: Examples of Chinese compound words with glosses of their constituent morphemes.

3.2. Chinese compound splitting

Although Chinese words exhibit far less surface variation than Arabic words do (Chinese has virtually no inflectional morphology), the incorporation of sub-lexical information is still advisable because of the high frequency of compound words in Chinese. While it is true that some compounds have idiosyncratic meanings that cannot be predicted from the meanings of the individual morphemes they are composed of, the relationship is often quite clear in other cases. Refer to to Figure 3 for a list of example compounds together with the meanings of their constituent morphemes.

To make use of this information, we rely on the insight that each Chinese character is a *logogram*, that is, it corresponds to a single morpheme. We therefore choose to adapt the smoothing method described in [12] to hierarchical phrase-based models. This involves extracting synchronous rules from both word- and character-segmented forms of the training data, splitting the word-segmented forms of the extracted rules into characters, combining both sets of rules, and finally renormalizing the probabilities.

surface	>y nwE mn Alnby∗ **sykwn** mnAsb l>smAk Alslmwn Alm$wyp .
segmented	Ay nwE mn Al nby∗ **swf ykwn** mnAsb l AsmAk Al slmwn Al m$wyp .
surface	>lA ywjd xTA fy AlfAtwrp ? AEtqd An **qymthA** >ElY mmA yjb >n tkwn Elyh .
segmented	AlA ywjd xTA fy Al fAtwrp ? AEtqd An **qymp hA** AElY mn hA yjb An tkwn ElY h .

Figure 2: Example of Arabic orthographic normalization and morpheme splitting.

Language	Condition	BLEU	OOV rate
Arabic-English	surface	44.75	6.0%
	+segmentation	45.45	3.7%
Chinese-English	surface	38.01	2.9%
	+segmentation	40.66	0.4%

Table 2: Impact of using sub-lexical information in model estimation for Arabic-English and Chinese-English translation in the DEV2 set (human transcription condition).

3.3. Experimental results

In this section, we present results for Arabic→English and Chinese→English translation, showing the benefits of using translation models that incorporate sub-lexical information. For Chinese, only the provided BTEC corpus was used in training. For Arabic, the BTEC corpus was analyzed using a tool based on the Buckwalter Morphological Analyzer. No additional training data were used. The model parameters were tuned to maximize BLEU score on a held-out development set (DEV3) as described in [13].

Table 2 shows the impact of the enhanced training procedure on the BLEU score and the OOV rate of the test sets used. For both language pairs, the increases in BLEU score and the reductions in OOV rate are statistically significant ($p < .05$).

Representative translations are shown in Table 3, showing that the BLEU score improvements correspond to improvements in translation quality. Note especially that the Chinese-English system was able to make use of the constituent elements of the three-character compound 便携式 to generate the hypothesis "*carry type*" instead of the more appropriate *portable*. Although unidiomatic, this translation preserves the most important aspects of the meaning in the sentence.

4. Punctuation restoration

In addition to the challenge of translating from imperfect recognizer output, spoken language translation faces the additional difficulty that punctuation is not present in the input but it is required in the output. Mauser et al. [15] review three strategies for predicting punctuation in the context of spoken language translation: preprocessing the source side to add punctuation and translating using a fully punctuated translation model, translating using a non-punctuated translation model and post-processing output to add punctuation, and translating using a model which implicitly adds punctuation during the translation process. While conceding that the first strategy of source preprocessing has several advantages, the authors decide in favor of the final option of implicit punctuation prediction to avoid the potential negative impact of falsely predicted source punctuation.

Since our translation system allows for the space of input hypotheses to be represented as a confusion network, we have the option of predicting source side punctuation but assigning only a limited probability to it, thus avoiding the pitfall Mauser et al. describes. In a flexible framework like ours, source-side punctuation prediction has the following advantages:

- The same translation model can be used for text and speech alike, since the source side will always contain punctuation.
- Punctuation prediction can make use of a richer feature set, including acoustic features from the speech signal (e.g., prosodic cues as to what the appropriate punctuation is). These features would not generally be available to a system that did implicit punctuation prediction.
- By utilizing a confusion network to assign probability to the presence of each punctuation character, the translation system can consider the input both with and without the punctuation, enabling longer phrases to be matched.

4.1. Punctuation prediction with an n-gram model

To predict source-side punctuation we used a trigram model trained on the punctuated source side of the training data. SRILM's `hidden-ngram` tool was used to insert periods, question marks, and commas so as to maximize the resulting probability of the sequences under the trigram model. For Arabic, end-of-sentence punctuation was moved to the start of the sentence since Arabic questions are frequently identifiable from words in the first position of the sentence. The predicted punctuation was then moved to the end of the sentence before translation. In Chinese, a language where question words remain *in situ*, no punctuation was dislocated.

4.2. Punctuation prediction in confusion networks

Instead of computing a punctuation hypothesis for every path in the input confusion network, we used the 1-best hypothe-

Language	Condition	Sample translation
Arabic-English	surface	Would you exchange it *lHjrp* overlooking the sea?
	segmented	Would you exchange it **for a room** overlooking the sea?
	surface	Could you *bglAf tglyfhA* for *bAlhdAyA*, please?
	segmented	Can you **wrap it up for cover** [ref. *giftwrap it*] for please?
Chinese-English	surface	I'd like 便携式 TV.
	+segmentation	I'd rather have a **carry type** [ref. *portable*] TV.
	surface	Where 营业部?
	+segmentation	Where is the **sales department**?

Table 3: Sample translations from the DEV2 set showing the effect of incorporating sub-lexical information in Chinese-Engish and Arabic-English translation.

Language pair	none	final-only	full
Chinese-English	15.31	16.55	17.72
Arabic-English	17.98	18.65	21.12

Table 4: Impact of source-side punctuation prediction on BLEU scores, DEV5 set, confusion network input.

sis from the confusion network (the so-called consensus transcription hypothesis), predicted punctuation for it, and then projected the predicted punctuation back on the source confusion network. A probability of 80% was assigned to each punctuation character; 20% was reserved for a null (epsilon) transition.[6]

4.3. Experimental results

Table 4 shows the effect of predicting punctuation in confusion networks on BLEU score. The table reports three conditions: no source-side punctuation, utterance-final punctuation only (periods and question marks), and full punctuation. Both sentence-final punctuation and sentence-internal punctuation have a large impact on translation quality. Figure 4 shows example translations.

5. The integrated system

Table 5 shows the performance of a baseline system using an unsmoothed surface translation model, the 1-best transcription hypothesis, and no source-side punctuation prediction as well as the performance of the integrated experimental system containing the three innovations described in this paper: source-side punctuation prediction, confusion network decoding for speech, and the incorporation of sub-lexical information in translation model estimation. The reported improvements are highly significant ($p < .01$) for both language pairs and evaluation metrics.

[6] As one reviewer correctly points out, there are a variety of ways of assigning probabilities to the predicted punctuation. Constant probabilities were used for simplicity, but more appropriate estimation methods are a matter of ongoing research.

Language	Condition	BLEU	TER
Arabic-English	surf, 1best, no-punc	15.50	70.45
	+seg, CN, full	21.09	58.61
Chinese-English	surf, 1best, no-punc	14.19	77.96
	+seg, CN, full	17.72	63.51

Table 5: Results of using sub-lexical information, confusion networks, and punctuation prediction on DEV5 ASR translation.

6. Conclusion

In this paper, we have presented the UMD hierarchical phrase-based statistical machine-translation system used in the IWSLT 2007 evaluation. We attempted to address three challenges of special interest to spoken language translation: 1) the paucity of training data, 2) the problem of imperfect recognition, and 3) the problem of missing punctuation.

The solutions described that deal with the problems of imperfect recognition and the problem of missing punctuation are language independent. Additionally, we have confirmed previous results that show an improvement for considering more than a 1-best transcription hypothesis, and we have argued for and demonstrated the benefit of carrying out source-side punctuation prediction before commencing translation. Future research will include more effective strategies for predicting punctuation, and more effective strategies for incorporating punctuation into confusion networks, such as those suggested in [16].

Finally, the technique we applied to mitigate the data sparseness problems, smoothing a "surface" translation model with sub-lexical information, has considerable potential. Although its execution in our experiments did rely on language-specific knowledge about word formation processes, the benefits for both Chinese and Arabic (languages with considerably different typologies) lead us to conclude that the basic smoothing technique should be generally applicable and likely to yield good results for many, if not all, language pairs. Future research will focus on automatic data-driven methods for determining appropriate morphological

Language	Condition	Sample translation
Arabic-English	none	Of course you can so please you could I try it in room it
	final-only	Of course you can so please you could I try it in room it.
	full	Of course, you can follow me, please, you can try it on in it.
	ref	Of course. You can follow me, please. You can try it on in the fitting room.
Chinese-English	none	No thank you may I'll
	final-only	No thank you may I'll.
	full	No, thank you, I can come by myself.
	ref	Oh. No thank you. I can manage myself.

Figure 4: Sample translations from DEV5 showing the effect of source-side punctuation prediction.

segmentation and analysis for an arbitrary source language.

7. Acknowledgements

The author would like to thank David Chiang for making the Hiero decoder sources available, Yejun Wu for his assistance preparing the Chinese examples, and two anonymous reviewers and especially Philip Resnik for their instructive comments. This work was supported by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001.

8. References

[1] Koehn, P., Och, F., and Marcu, D., "Statistical phrase based translation," in *Proc. of HLT-NAACL '03*, 2003.

[2] Chiang, D., "Hierarchical phrase-based machine translation," *Computational Linguistics* 33(2):201–228, 2007.

[3] Zollmann, A., and Venugopal, A., "Syntax augmented machine translation via chart parsing," in *Proc. of the SMT Workshop, HLT-NAACL 2006*, New York, 2006.

[4] Ney, H. "Speech translation: Coupling of recognition and translation," in *Proc. of IEEE Int. Conference on Acoustic, Speech and Signal Processing*, pp. 517–520, 1999.

[5] Bertoldi, N., Zens, R., and Federico, M., "Speech translation by confusion network decoding," in *Proc. of the Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[6] Cheppalier, J., and Rajman, M., "A generalized CYK algorithm for parsing stochastic CFG," in *Proc. of the Workshop on Tabulation in Parsing and Deduction (TAPD98)*, pp. 133–137, 1998.

[7] Dyer, C., and Resnik, P. "Word lattice parsing for statistical machine translation," Technical report, University of Maryland, College Park, 2007.

[8] Mangu, L., Brill, E., and Stolcke, A., "Finding consensus in speeh recognition: Word error minimization and other applications of confusion networks," *Speech and Language* 14(4):373–400, 2000.

[9] Vilar, D., Peter, J.-T., and Ney, H., "Can we translate letters?," in *Proc. of the Second Workshop on Statistical Machine Translation*, pp. 33–39, 2007.

[10] Lee, Y.-S., "Morphological analysis for statistical machine translation," in *Proc. of HLT-NAACL 2004: Companion Volume*.

[11] Habash, N., and Sadat, F., "Arabic preprocessing schemes for statistical machine translation," in *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL) 2006*.

[12] Shen, W., Zens, R., Bertoldi, N., and Federico, M., "The JHU workshop 2006 IWSLT system," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT) 2006*, pp. 59–63, 2006.

[13] Och, F.-J., and Ney, H., "Discriminitive training and maximum entropy models for statistical machine translation," in *Proc. of the 40th Annual Meeting of the ACL*, pp. 295–302, 2002.

[14] Koehn, P., "Statistical significance tests for machine translation evaluation," in *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 388–395, 2004.

[15] Mauser, A., Zens, R., Matusov, E., Hasan, S., and Ney, H., "The RWTH statistical machine translation system for the IWSLT 2006 evaluation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT) 2006*, pp. 103–110, 2006.

[16] Cattoni, R., Bertoldi, N., and Federico, M., "Punctuating confusion networks for speech translation", in *Proc. of Interspeech*, 2007.