Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation

Andreas Zollmann and **Ashish Venugopal** and **Stephan Vogel**
School of Computer Science
Carnegie Mellon University
{zollmann,ashishv,stephan.vogel}@cs.cmu.edu

Abstract

Statistical machine translation (SMT) is based on the ability to effectively learn word and phrase relationships from parallel corpora, a process which is considerably more difficult when the extent of morphological expression differs significant across the source and target languages. We present techniques that select appropriate word segmentations in the morphologically rich source language based on contextual relationships in the target language. Our results take advantage of existing word level morphological analysis components to improve translation quality above state-of-the-art on a limited-data Arabic to English speech translation task.

1 Introduction

The problem of translating from a language exhibiting rich inflectional morphology to a language exhibiting relatively poor inflectional morphology presents several challenges to the existing components of the statistical machine translation (SMT) process. This inflection gap causes an abundance of surface word forms [1] in the source language compared with relatively few forms in the target language. This mismatch aggravates several issues found in natural language processing: more unknown words forms in unseen data, more words occurring only once, more distinct words and lower token-to-type ratios (mean number of occurrences over all distinct words) in the source language than in the target language.

Lexical relationships under the standard IBM models (Brown et al., 1993) do not account for many-to-many mappings, and phrase extraction relies heavily on the accuracy of the IBM word-to-word alignment. In this work, we propose an approach to bridge the inflectional gap that addresses the issues described above through a series of preprocessing steps based on the Buckwalter Arabic Morphological Analyzer (BAMA) tool (Buckwalter, 2004). While (Lee et al., 2003) develop accurate segmentation models of Arabic surface word forms using manually segmented data, we rely instead on the translated context in the target language, leveraging the manually constructed lexical gloss from BAMA to select the appropriate segmented sense for each Arabic source word.

Our technique, applied as preprocessing to the source corpus, splits and normalizes surface words based on the target sentence context. In contrast to (Popovic and Ney, 2004) and (Nießen and Ney, 2004), we do not modify the IBM models, and we leave reordering effects to the decoder. Statistically significant improvements (Zhang and Vogel, 2004) in BLEU and NIST translation score over a lightly stemmed baseline are reported on the available and well known BTEC IWSLT'05 Arabic-English corpus (Eck and Hori, 2005).

[1] We use the term surface form to refer to a series of characters separated by whitespace

2 Arabic Morphology in Recent Work

Arabic-to-English machine translation exemplifies some of the issues caused by the inflection gap. Refer to (Buckwalter, 2005) and (Larkey et al., 2002) for examples that highlight morphological inflection for a simple Modern Standard Arabic (MSA) word and basic stemming operations that we use as our baseline system.

(Nießen and Ney, 2000) tackle the inflection gap for German-to-English word alignment by performing a series of morphological operations on the German text. They fragment words based on a full morphological analysis of the sentence, but need to use domain specific and hand written rules to deal with ambiguous fragmentation. (Nießen and Ney, 2004) also extend the corpus by annotating each source word with morphological information and building a hierarchical lexicon. The experimental results show dramatic improvements from sentence-level restructuring (question inversion, separated verb prefixes and merging phrases), but limited improvement from the hierarchical lexicon, especially as the size of the training data increases.

We conduct our morphological analysis at the word level, using Buckwalter Arabic Morphological Analyzer (BAMA) version 2.0 (Buckwalter, 2004). BAMA analyzes a given surface word, returning set of potential segmentations (order of a dozen) of that word into prefixes, stems, and suffixes. Our techniques select the appropriate splitting from that set by taking into account the target sides (full sentences) of that word's occurrences in the training corpus. We now describe each splitting techniques that we apply.

2.1 BAMA: Simple fragment splitting

We begin by simply replacing each Arabic word with the fragments representing the first of the possible splittings returned by the BAMA tool. BAMA uses simple word-based heuristics to rank the splitting alternatives.

2.2 CONTEXT: Single Sense selection

In the step CONTEXT, we take advantage of the gloss information provided in BAMA's lexicon. Each potential splitting corresponds to a particular choice of prefix, stem and suffix, all of which exist in the manually constructed lexicon, along with a set of possible translations (*glosses*) for each fragment. We select fragmentations whose corresponding glosses have the most target side matches in the parallel translation. The choice of fragmentation is saved and used for all occurrences of the surface form word in training and testing, introducing context sensitivity without parsing solutions. In case of an unseen word occurring during testing, we segment it simply using the first alternative from the BAMA tool. This allows us to still translate an unseen test word correctly even if the individual fragments of that segmentation were never seen during training.

2.3 CORRMATCH: Correspondence matching

The Arabic language often encodes linguistic information within the surface word form that is not present in English. Word fragments that represent this missing information are misleading in the translation process unless explicitly aligned to the NULL word on the target side. In this step we explicitly remove fragments that correspond to lexical information that is not represented on the English side. While (Lee, 2004) builds part of speech models to recognize such elements, we use the fact that their corresponding English translations in the BAMA lexicon are empty. Examples of such fragments are case and gender markers.

3 Experimental Framework

We evaluate the impact of inflectional splitting on the BTEC (Takezawa et al., 2002) IWSLT05 Arabic language data track. The "Supplied" data track includes a 20K Arabic/English sentence pair training set, as well as a development ("DevSet") and test ("Test05") set of 500 Arabic sentences each and 16 reference translations per Arabic sentence. Details regarding the IWSLT evaluation criteria and data topic and collection methods are available in (Eck and Hori, 2005). We also evaluate on test and development data randomly sampled from the complete supplied dev and test data, due to considerations noted by (Josep M.Crego, 2005) regarding the similarity of the development and test data sets.

Inflection system	Source Text	Translation output
No preprocessing	UTF8 Arabic	this is still change clean
Orthographic normalization (baseline)	UTF8 Arabic	this is still change clean
BAMA	h‘*A lA ya zAl u gayor naZiyf	this still to cleaned
BAMA+CONTEXT+CORRMATCH	h‘*A lA ya zAl _ gayor naZiyf	this is still others cleaned
Reference translation	NA	this still isn’t clean

Table 1: Translation example (test set) at each successive stage of inflectional splitting. Note the removal of the ”u” (underlined) when using the CORRMATCH setting

3.1 System description

Translation experiments were conducted using the (Vogel et al., 2003) system with reordering and future cost estimation. We trained translation parameters for 10 scores (language model, word and phrase count, and 6 translation model scores from (Vogel, 2005)) with Minimum Error Rate training on the development set. We optimized separately for both the NIST (Doddington, 2002) and the BLEU metrics (Papineni et al., 2002).

4 Translation Results

Table 2 and 3 shows the results of each stage of inflectional splitting on the BLEU and NIST metrics. Basic orthographic normalization serves as a baseline (merging all Alif forms to the base form). The test set NIST scores in Table 2 shows steady improvements up to 5 percent relative, as more sophisticated splitting techniques are used up to BAMA+CONTEXT+CORRMATCH. These improvements are statistically significant over the baseline in both metrics.

Our NIST results for all the final stages of inflectional splitting would place us above the top NIST scores from the ISWLT evaluation on the supplied test set.[2] On both DevSet/Test05 and the randomly split data, we see more dramatic improvements in the NIST scores than in BLEU. This might be due to the NIST metric’s sensitivity to correctly translating certain high gain words in the test corpus. Inflectional splitting techniques that cause previously unknown surface form words to be translated correctly after splitting can significantly impact the overall score.

5 Conclusion and Future Work

This work shows the potential for significant improvements in machine translation quality by directly bridging the inflectional gap across language pairs. Our method takes advantage of source and target language context when conducting morphological analysis of each surface word form, while avoiding complex parsing engines or refinements to the alignment training process. Our results are presented on moderately sized corpora rather than the scarce resource domain that has been traditionally employed to highlight the impact of detailed morphological analysis.

By showing the impact of simple processing steps we encourage the creation of simple word and gloss level analysis tools for new languages and show that small investments in this direction (compared to high octane context sensitive parsing tools) can yield dramatic improvements, especially when rapid development of machine translation tools becomes increasingly relevant to the research community. While our work focused on processing the morphologically rich language and then translating ”down” into the morphologically poor language, we plan to use the analysis tools developed here to model the reverse translation process as well, the harder task of translating ”up” into a highly inflected space.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC Catalog No. LDC2004L02, Linguistic Data Consortium, www.ldc.upenn.edu/Catalog.

[2] The IWSLT evaluation did not allow systems to train separately for evaluation on BLEU or NIST, but results from the proceedings indicate that top performers in each metric optimized towards the respective metric.

Inflection system	NIST – Dev.	**NIST – Test**	BLEU – Dev.	**BLEU – Test**
No preprocessing	9.33	9.44	51.1	49.7
Orthographic normalization (baseline)	9.41	9.51	51.0	49.7
BAMA	9.90	9.76 (+2.5%)	52.0	50.2 (+1%)
BAMA+CONTEXT+CORRMATCH	9.91	**10.02** (+5.3%)	52.8	**52.0** (+4.7%)

Table 2: Translation results for each stage of inflectional splitting for the merged, sampled dev. and test data, highest scores in bold, relative improvements in brackets

Inflection system	NIST – Dev.	**NIST – Test**	BLEU – Dev.	**BLEU – Test**
No preprocessing	9.46	9.38	51.1	49.6
Orthographic normalization (baseline)	9.58	9.35	52.1	49.8
BAMA	10.10	9.60 (+2.7%)	53.8	48.8 (-2%)
BAMA+CONTEXT+CORRMATCH	10.08	**9.79** (+4.7%)	53.7	**50.6** (+1.6%)

Table 3: Translation results for each stage of inflectional splitting for the BTEC Supplied DevSet/Test05 data, highest scores in bold, relative improvements in brackets

Tim Buckwalter. 2005. www.qamus.org/morphology.htm.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *In Proc. ARPA Workshop on Human Language Technology*.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation*, pages 11–17.

Jose B.Marino Josep M.Crego, Adria de Gispert. 2005. The talp ngram-based smt system for iwslt'05. In *Proceedings of International Workshop on Spoken Language Translation*, pages 191–198.

Leah Larkey, Lisa Ballesteros, and Margaret Connell. 2002. Improving stemming for arabic information retrieval: Light stemming and co-occurrence analysis. In *Proc. of the 25th annual international ACM SIGIR conference on Research and development information retrieval*.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language model based arabic word segmentation. In *ACL*, Sapporo, Japan, July 6-7.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Boston,MA, May 27-June 1.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *The 18th International Conference on Computational Linguistics*.

Sonja Nießen and Herman Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Comput. Linguist.*, 30(2):181–204.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics*, pages 311–318.

H. Popovic and Hermann Ney. 2004. Improving word alignment quality using morpho-syntactic information. In *20th International Conference on Computational Linguistics (CoLing), Geneva, Switzerland*.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of LREC 2002*, pages 147–152, Las Palmas, Canary Islands, Spain, May.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venogupal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical translation system. In *Proceedings of MT Summit IX*, New Orleans, LA, September.

Stephan Vogel. 2005. PESA: Phrase pair extraction as sentence splitting. In *Proceedings of MT Summit X*, Phuket,Thailand, September.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMII)*, Baltimore, MD, October.