Morphological Modeling for Machine Translation of English-Iraqi Arabic Spoken Dialogs

Katrin Kirchhoff
Department of
Electrical Engineering
University of Washington
Seattle, WA, USA
kk2@u.washington.edu

Wilson Tam
Microsoft Corporation
wilson.yctam@gmail.com

Colleen Richey, Wen Wang
SRI International
Menlo Park, CA, USA
colleen@speech.sri.com
wwang@speech.sri.com

Abstract

This paper addresses the problem of morphological modeling in statistical speech-to-speech translation for English to Iraqi Arabic. An analysis of user data from a real-time MT-based dialog system showed that generating correct verbal inflections is a key problem for this language pair. We approach this problem by enriching the training data with morphological information derived from source-side dependency parses. We analyze the performance of several parsers as well as the effect on different types of translation models. Our method achieves an improvement of more than a full BLEU point and a significant increase in verbal inflection accuracy; at the same time, it is computationally inexpensive and does not rely on target-language linguistic tools.

1 Introduction

SMT from a morphologically poor language like English into a language with richer morphology continues to be a problem, in particular when training data is sparse and/or the SMT system has insufficient modeling capabilities for morphological variation in the target language. Most previous approaches to this problem have utilized a translate-and-inflect method, where a first-pass SMT system is trained on lemmatized forms, and the correct inflection for every word is predicted in a second pass by statistical classifiers trained on a combination of source and target language features. This paper looks at morphological modeling from a different perspective, namely to improve SMT in a real-time speech-to-speech translation system. Our focus is on resolving those morphological translation errors that are most likely to cause confusions and misunderstandings in machine-translation mediated human-human dialogs. Due to the constraints imposed by a real-time system, previous approaches that rely on elaborate feature sets and multi-pass processing strategies are unsuitable for this problem. The language pair of interest in this study is English and Iraqi Arabic (IA). The latter is a spoken dialect of Arabic with few existing linguistic resources. We therefore develop a low-resource approach that relies on source-side dependency parses only. We analyze its performance in combination with different types of parsers and different translation models. Results show a significant improvement in translation performance in both automatic and manual evaluations. Moreover, the proposed method is sufficiently fast for a real-time system.

2 Prior Work

Much work in SMT has addressed the issue of translating from morphologically-rich languages by preprocessing the source and/or target data by e.g., stemming and morphological decomposition (Popovic and Ney, 2004; Goldwater and McClosky, 2005), compound splitting (Koehn and Knight, 2003), or various forms of tokenization (Lee, 2004; Habash and Sadat, 2006). In (Minkov et al., 2007; Toutanova et al., 2008) morphological generation was applied as a postprocessing step for translation *into* morphologically-rich languages. A maximum-entropy Markov model was trained to predict the correct inflection for every stemmed word in the

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 995–1000,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

machine translation output from a first-pass system, conditioned on a set of lexical, morphological and syntactic features. More recently, (Chahuneau et al., 2013) applied a similar translate-and-inflect approach, utilizing unsupervised in addition to supervised morphological analyses. Inflection generation models were also used by (Fraser et al., 2012; Weller et al., 2013) for translation into German, and by (El Kholy and Habash, 2012) for Modern Standard Arabic. (Sultan, 2011) added both syntactic information on the source side that was used in filtering the phrase table, plus postprocessing on the target side for English-Arabic translation. Still other approaches enrich the translation system with morphology-aware feature functions or specific agreement models (Koehn and Hoang, 2007; Green and DeNero, 2012; Williams and Koehn, 2011).

In contrast to the above studies, which have concentrated on text translation, this paper focuses on spoken language translation within a bilingual human-human dialog system. Thus, our main goal is not to predict the correct morphological form of every word, but to prevent communication errors resulting from the mishandling of morphology. The intended use in a real-time dialog system imposes additional constraints on morphological modeling: any proposed approach should not add a significant computational burden to the overall system that might result in delays in translation or response generation. Our goal is also complicated by the fact that our target language is a spoken dialect of Arabic, for which few linguistic resources (training data, lexicons, morphological analyzers) exist. Lastly, Arabic written forms are morphologically highly ambiguous due to the lack of short vowel markers that signal grammatical categories.

3 Dialog System and Analysis

The first step in the dialog system used for this study consists of an automatic speech recognition (ASR) component that produces ASR hypotheses for the user's speech input. Several error detection modules then identify likely out-of-vocabulary and misrecognized words. This information is used by a clarification module that asks the user to rephrase these error segments; another module then combines the user's answers into a merged, corrected representation before sending it to the translation engine. A machine translation error detection module analyzes the translation to check for errors, such as unknown words. If an error is found, another clarification subdialog is initiated; otherwise, the translation is sent to a text-to-speech engine to produce the acoustic output in the other language. A schematic representation is shown in Figure 1. More details about the system can be found in (et al., 2013). The system was evaluated in live mode with native IA speakers as part of the DARPA BOLT Phase-II benchmark evaluations. The predefined scenarios included military and humanitarian assistance/disaster relief scenarios as well as general topics. All system interactions were logged and evaluated by bilingual human assessors.

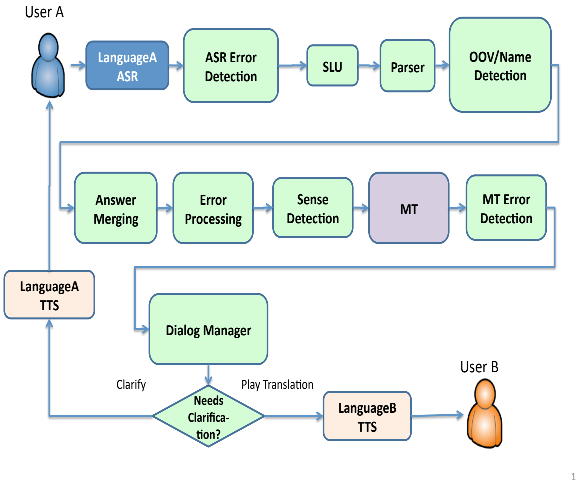

Figure 1: Dialog system used in this work.

During debriefing sessions with the users, some users voiced dissatisfaction with the translation quality, and a subsequent detailed error analysis was conducted on the logs of 30 interactions. Similar to previous studies (Condon et al., 2010) we found that a frequently recurring problem was wrong morphological verb forms in the IA output. Some examples are shown in Table 1. In Example 1, *to make sure* should be translated by a first-person plural verb but it is translated by a second-person plural form, changing the meaning to (*you (pl.) make sure*). The desired verb form would be *ntAkd*. Similarly, in Example 2 the translation of *transport* should agree with the translations of *someone* and the preceding

1	*you need to tell the locals to evacuate the area so we can secure the area* **to make sure** *no one gets hurt* *lAzm tqwl Alhm AhAly AlmnTqp bAlAxlA' AlmnTqp HtY nqdr nwmn AlmnTqp Elmwd* **ttAkdwn** *Anh mHd ytAY*
2	*do you have someone that can* **transport you** *to the nearest american base* *Endk wAHd yqdr* **nqlk** *lAqrb qAEdp Amrykyp*

Table 1: Examples of mistranslated morphology: English ASR hypotheses and IA translation hypotheses.

auxiliary verb *can* (*yqdr*). The correct form would be *yqlk* (*he/she transports you*) instead of *nqlk* (*we transport you*). Such translation errors are confusing to users as they affect the understanding of basic semantic roles. They tend to occur when translating English infinitival constructions (*to*+verb) or other syntactic constructions where English base verb forms need to be translated by a finite verb in IA. In these cases, explicit morphological features like person and number are required in Arabic but they are lacking in the English input.

4 Approach

An analysis of the SMT component showed that morphological translation errors primarily occur when a head word and its dependent (such as a verbal head and its subject noun dependent) are translated as part of different phrases or rules. In that case, insufficient context is available to produce the correct translation. Our approach is to annotate syntactic dependencies on the source side using a statistical parser. Based on the resulting dependency structures the source-side data is then tagged with explicit morphological verbal features using deterministic rules (e.g., subject nouns assign their person/number features to their verbal heads), and a new translation model is trained on this data. Our assumption is that words tagged with explicit morphological features will be aligned with their correct translations during training and will thus produce correctly inflected forms during testing even when the syntactic context is not available in the same phrase/rule. For instance, the input sentence in Example 1 in Table 1 would be annotated as:
you need-2sg to tell-2sg the locals to evacuate-3pl the area so we can-1pl secure-1pl the area to make-1pl sure no one gets-3sg hurt.
This approach avoids the costly extraction of multiple features, subsequent statistical classification, and inflection generation during run time; moreover, it does not require target-side annotation tools, an advantage when dealing with under-resourced spoken dialects. There are, however, several potential issues with this approach. First, introducing tags fragments the training data: the same word may receive multiple different tags, either due to genuine ambiguity or because of parser errors. As a result, word alignment and phrase extraction may suffer from data sparsity. Second, new word-tag combinations in the test data that were not observed in the training data will not have an existing translation. Third, the performance of the model is highly dependent on the accuracy of the parser. Finally, we make the assumption that the expression of person and number categories are matched across source and target language – in practice, we have indeed seen very few mismatched cases where e.g., a singular noun phrase in English is translated by a plural noun phrase in IA (see Section 6 below).

To address the first point the morph-tagged translation model can be used in a backoff procedure rather than as an alternative model. In this case the baseline model is used by default, and the morph-tagged model is only used whenever heads and dependents are translated as part of different phrases. Unseen translations for particular word-tag combinations in the test set could in principle be addressed by using a morphological analyzer to generate novel word forms with the desired inflections. However, this would require identifying the correct stem for the word in question, generating all possible morphological forms, and either selecting one or providing all options to the SMT system, which again increases system load. We analyzed unseen word-tag combination in the test data but found that their percentage was very small (< 1%). Thus, for these forms we back off to the untagged counterparts rather than generating new inflected forms. To obtain better insight into the effect of parsing accuracy we compared the performance of two parsers in our annotation pipeline: the Stanford parser (de Marneffe et al., 2006) (version 3.3.1) and the Macaon parser (Nasr et al., 2014). The latter is an implementation of graph-based parsing (McDonald et al., 2005) where a projective dependency tree maximizing a score function is sought in the graph of all possible trees using dynamic programming. It uses a 1st-order decoder, which is more robust to speech input as well as out-of-domain training data. The features implemented reflect those of (Bohnet, 2010) (based on lexemes and part-of-speech tags). The parser was trained on Penn-Treebank data transformed to match speech (lower-cased, no punctuation), with one iteration of self-training on the Transtac training set. We also use the combination of both parsers, where source words are only tagged if the tags derived independently from each parser agree with each other.

5 Data and Baseline Systems

Development experiments were carried out on the Transtac corpus of dialogs in the military and medical domain. The number of sentence pairs is 762k for the training set, 6.9k for the dev set, 2.8k for eval set 1, and 1.8k for eval set 2. Eval set 1 has one reference per sentence, eval set 2 has four references. For the development experiments we used a phrase-based Moses SMT system with a hierarchical reordering model, tested on Eval set 1. The language model was a backoff 6-gram model trained using Kneser-Ney discounting and interpolation of higher- and lower-order n-grams. In addition to automatic evaluation we performed manual analyses of the accuracy of verbal features in the IA translations on a subset of 65 sentences (containing 143 verb forms) from the live evaluations described above. This analysis counts a verb form as correct if its morphological features for person and number are correct, although it may have the wrong lemma (e.g., wrong word sense). The development experiments were designed to identify the setup that produces the highest verbal inflection accuracy. For final testing we used a more advanced SMT engine on Eval set 2.This system is the one used in the real-time dialog system; it contains a hierarchical phrase-based translation model, sparse features, and a neural network joint model (NNJM) (Devlin et al., 2014).

	BLEU		Acc (%)	
Parser	std	bo	std	bo
Baseline	16.8	N/A	37.1	N/A
Stanford	16.9	17.0	60.1	59.4
Macaon	17.0	17.1	**67.1**	62.9
Combined	17.1	17.1	59.4	57.3

Table 2: BLEU scores on Transtac eval set 1 and accuracy of verbal morphological features on manual eval set. *std* = standard, *bo* = backed-off system.

6 Experiments and Results

Results in Table 2 show the comparison between the baseline, different parsers, and the combined system. We see that verbal inflection accuracy increases substantially from the baseline performance and is best for the Macaon parser. Improvements over the baseline system without morphology are statistically significant; differences between the individual parsers are not (not, however, that the sample size for manual evaluation was quite small).

BLEU is not affected negatively but even increases slightly - thus, data fragmentation does not seem to be a problem overall. This may be due to the nature of the task and domain, which is results in fairly short, simple sentence constructions that can be adequately translated by a concatenation of shorter phrases rather than requiring longer phrases. Back-off systems (indicated by *bo*) and the combined system improve BLEU only trivially while decreasing verbal inflection accuracy by varying amounts. For testing within the dialog system we thus choose the Macaon parser and utilize a standard translation model rather than a backoff model. An added benefit is that the Macaon parser is already used in other components in the dialog system. Using this setup we ran two experiments with dialog system's SMT engine: first, we re-extracted phrases and rules based on the morph-tagged data and reoptimized the feature weights. In the second experiment, we additionally applied the NNJM to the morph-tagged source text. To this end we include all the morphological variants of the original vocabulary that was used for the NNJM in the untagged baseline system. Table 3 shows the results. The morph-tagged data improves the BLEU score under both conditions: in Experiment 1, the improvement is almost a full BLEU point (0.91); in Experiment 2 the improvement is even larger (1.13), even though the baseline performance is stronger. Both results are statistically significant at p = 0.05, using a paired bootstrap resampling test. The combination of morph-tagged data and the more advanced modeling options (sparse features, NNJM) in this system seem to be beneficial. Improved translation performance may also be captured by the four reference translations as opposed to one in Eval set 1. In order to assess the added computation cost of our procedure we computed the decoding speed of the MT component in the dialog system for both the baseline and the morpho-tag systems. In the baseline MT system (with NNJM) without morpho-tags, decoding takes 0.01572 seconds per word or 0.15408 seconds per sentence – these numbers were obtained on a Dell Precision M4800 Laptop with a quad-core Intel i7-4930MX Processor and 32GB of RAM. Morpho-tagging only adds 0.00031 seconds per word or 0.0024 seconds per sentence. Thus, our procedure is extremely efficient.

System	no NNJM	with NNJM
Baseline	34.38	36.17
Morph tags	**35.29**	**37.30**

Table 3: BLEU on Eval set 2 using dialog system's SMT engine.

An analysis of the remaining morphological translation errors not captured by our approach showed that in about 34% of all cases these were due to part-of-speech tagging or parser errors, i.e. verbs were mistagged as nouns rather than verbs and thus did not receive any morphological tags, or the parser hypothesized wrong dependency relations. In 53% of the cases the problem is the lack of more extensive discourse or contextual knowledge. This includes constructions where there is no overt subject for a verb in the current utterance, and the appropriate underlying subject must be inferred from the preceding discourse or from knowledge of the situational context. This is an instance of the more general problem of control (see e.g.,(Landau, 2013) for an overview of research in this area). It is exemplified by cases such as the following:

1. *The first step is to make sure that all personnel are in your debrief.*

Here, the underlying subject of "to make sure" could be a range of different candidates (*I*, *you*, *we*, etc.) and must be inferred from context.

2. *I can provide up to one platoon to help you guys cordon off the area.*

In this case the statistical parser identified *I* as the subject of *help*, but *platoon* is more likely to be the controller and was in fact identified as the underlying subject by the annotator. Such cases could potentially be resolved during the parsing step by integrating semantic information, e.g. as in (Bansal et al., 2014). However, initial investigations with semantic features in the Macaon parser resulted in a significant slow-down of the parser. In other cases, more sophisticated modeling of the entities and their relationships in the situational context will be required. This clearly is an area for future study.

Finally, in 13% of the cases, mistranslations are caused by a mismatch of number features across languages (e.g. number features for nouns such as *family* or *people*).

7 Conclusion

We have shown that significant gains in BLEU and verbal inflection accuracy in speech-to-speech translation for English-IA can be achieved by incorporating morphological tags derived from dependency parse information in the source language. The proposed method is fast, low-resource, and can easily be incorporated into a real-time dialog system. It adds negligible computational cost and does not require any target-language specific annotation tools. Possible areas for future study include the use of discourse or and other contextual information to determine morphological agreement, application to other languages pairs/morphological agreement types, and learning the annotation rules from data.

Acknowledgments

This study was funded by the Defense Advanced Research Projects Agency (DARPA) under contract HR0011-12-C-0016 - subcontract 19-000234.

References

M. Bansal, K. Gimpel, and K. Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proc. of ACL*.

B. Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*, pages 89–97.

V. Chahuneau, E. Schlinger, N. Smith, and C. Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of EMNLP*.

S. Condon, D. Parvaz, J. Aberdeen, C. Doran, A. Freeman, and M. Awad. 2010. Evaluation of machine translation errors in English and Iraqi Arabic. In *Proceedings of LREC*.

M.-C. de Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

J. Devlin et al. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of ACL*, pages 1370–1380.

A. El Kholy and N. Habash. 2012. Translate, predict or generate: modeling rich morphology in statistical machine translation. In *Proceedings of EAMT*.

N.F. Ayan et al. 2013. Can you give me another word for hyperbaric? - improving speech translation using targeted clarification questions. In *Proceedings of ICASSP*.

A. Fraser, M. Weller, A. Cahill, and F. Cap. 2012. Modeling inflection and word-formation in SMT. In *Proceedings of EACL*, pages 664–674.

S. Goldwater and D. McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of EMNLP*, pages 676–683.

S. Green and J. DeNero. 2012. A class-based agreement model for generating accuractely inflected translation. In *Proceedings of ACL*, pages 146–155.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of NAACL*.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proceedings of EMNLP*, pages 868–876.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *Proceedings of EACL*.

I. Landau. 2013. *Control in Generative Grammar: A Research Companion*. Cambridge University Press, Cambridge, UK.

Y.S. Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL*.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP*, pages 523–530.

E. Minkov, K. Toutanova, and H. Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of ACL*, pages 128–135.

A. Nasr, F. Bechet, B. Favre, T. Bazillon, J. Deulofeu, and A. Valli. 2014. Automatically enriching spoken corpora with syntactic information for linguistic studies. In *Proceedings of LREC*.

M. Popovic and H. Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of LREC*.

S. Sultan. 2011. *Applying Morphology to English-Arabic Statistical Machine Translation*. Ph.D. thesis, Department of Computer Science, ETH Zürich.

K. Toutanova, H. Suzuki, and A. Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL*, pages 514–522.

M. Weller, A. Fraser, and S. Schulte im Walde. 2013. Using subcategorization knowledge to improve case prediction for translation to German. In *Proceedings of ACL*, pages 593–603.

P. Williams and P. Koehn. 2011. Agreement constraints for statistical machine translation into German. In *Proceedings of WMT*.