interspeech 2008 incorporating SST 08

Strategies for Building a Farsi-English SMT System from Limited Resources

Andreas Kathol, Jing Zheng

SRI International,
Speech Technology and Research Lab
Menlo Park, CA, 94025
{kathol,zj}@speech.sri.com

Abstract

One of the recent tasks for machine translation research has been development of translation capabilities in a time frame as short as 100 days. Such a task requires developers to consider what can be done with relatively small amounts of data in a small time frame. This inherently limits the type and complexity of the effort to be devoted to this task. In this paper we will focus on the kinds of improvements for a Farsi-to-English translation system achieved by means of algorithmic changes, adding raw, domain-unspecific resources, and unsupervised morphological segmentation. The cumulative effect of these measures has been an improvement in BLEU scores of about 25% relative on an internal test set.

Index Terms: statistical machine translation, low resource languages

1. Introduction

The advent of purely data-driven, statistical methods in machine translation has made it possible to extract the human knowledge implicit in large bodies of multilingual texts. Unlike earlier purely knowledge-based approaches, which required the painstaking adoption of linguistic knowledge in the form of explict rules, statistical machine translation (SMT) allows for the development of a working system in a relatively short amount of time. Since the amount of available training data plays a key role in determining the performance of a statistically based system, it is generally understood that the best way to improve a system's performance is to supply more training data.

For a given language pair, however, the amount of overall available resources might be quite limited; hence, performance improvement will have to be sought from other sources. One of the tasks in the current DARPA TransTac program is the rapid development of translation capabilities in a time frame as short as 100 days. Such a task requires developers to consider what can be done with relatively small amounts of data in a small time frame. This inherently limits the type and complexity of the effort to be devoted to this task. For instance, given the available resources of the language in question, it might not be practical to develop a reliable morphological analyzer for a particular language in only three months, let alone a rule-based translation engine such as the one for English to Iraqi Arabic employed in SRI's IraqComm system ([1]).

A case in point is Farsi, which was the surprise language chosen for the 2007 TransTac task of developing translation capabilities in at most 100 days. While SRI did not participate in the evaluation itself, we took the opportunity to study the effect of various methods for improving the performance of our SMT system. In this paper we will focus on the kinds of improvements achieved by means of (1) algorithmic changes, (2) adding domain-unspecific resources, and (3) unsupervised morphological segmentation. In particular, we were interested in methods to improve translation quality that can be brought to bear quickly and easily without any language-specific knowledge. While the task encompasses all aspects of spoken-language translation, we will concentrate here on the text-to-text part only.

	Train	Dev	Internal Test
Lines	75870	1380	4736
EN word tokens	580958	9235	34973
EN word types	10777	1654	2872
FA word tokens	497558	10754	31546
FA word types	23401	2882	5184

Table 1: Farsi data quantities

2. Approach

We describe a number of experiments conducted on the TransTac Farsi data. After discussing the various initial preprocessing steps applied to the data, we describe SRI's statistical machine translation system SRInterp™. We subsequently present results of the baseline system followed by various extensions.

2.1. Data

The basis for the Farsi SMT system was the data supplied by DARPA for the 2007 surprise language evaluation. From this initial set 85,400 nonempty aligned sentence pairs were set aside for training. As an initial data cleaning step, pairs containing ASR fragments (such as *we tr- we try to provide ...*) or ASR "reject" symbols were eliminated. A summary of the resulting data quantities is given in Table 1.

The next preprocessing step consisted of eliminating filled pauses (e.g., *%um*), miscellaneous markup (e.g., *%breath*) and punctuation symbols from the data. In addition, some of the dialectal and orthographic variation in the Farsi data was normalized. For instance, 'his.name OBJ' occurs both in the conventional spelling *AsmS rA* or as *AsmSv*.[1] To this end, a set of 12,561 normalization mappings supplied by NIST was applied to the data, which led to an almost 20% reduction of the vocab-

[1] All Farsi examples are given here in USCPers transliteration, see [2].

	Orig	with normalization
Tokens	497558	514872
Types	23402	18871

Table 2: Effect of normalization on Farsi training data

	BLEU score
Internal Test	0.289

Table 3: Baseline BLEU scores for standard phrase-based Farsi-to-English SMT system

ulary size of the training set, as shown in Table 2.

A corresponding set of 199 replacement rules was applied to the English side, regularizing expressions such as *don't* to *do not*, but yielding only a 0.5% reduction in vocabulary size.

Finally, the Farsi data was transliterated from Persian script to a purely ASCII-based format using the USCPers transliteration scheme, cf. [2].

2.2. SRInterp statistical machine translation system

The SRInterp engine is SRI's SMT decoder ([3]), which supports both standard phrase-based ([4]) and hierarchical phrase-based translation methods ([5], [6]). The standard phrase-based translation is based on a bilingual phrase-pair translation model. Compared to earlier methods based on word-for-word translation, phrase-based approaches are superior at memorizing training data and are better at modeling local word reordering. The standard phrase-based approach, however, cannot directly model correspondences that involve long-distance relationships. Such correspondences can pose serious problems for language pairs with rather different word orders. By contrast, hierarchical phrase-based translation is based on lexicalized synchronous context-free grammars that are far superior at modeling long-distance dependencies and hence provide a more principled approach to word reordering models. The improved ability to deal with word order mismatches seems to be borne out in the difference in performance by the phrase-based vs. the hierarchical variants on the Farsi/English data as shown in the next sections.

2.3. Baseline standard phrase-based system

We first used the data described in Section 2.1 to train a standard phrase-based system and evaluated it against both the internal test set. The performance of this baseline system as measured in BLEU scores ([7]) is given in Table 3.

3. Improvements

Taking the standard phrase-based system as the point of departure, we investigated strategies of improving the SMT performance. These consisted of (1) changing the underlying translation approach, (2) adding freely available but domain-unspecific resources, and (3) applying morphological segmentation.

3.1. Hierarchical phrase-based system

The same data was used to train a hierarchical phrase-based SMT system. We observed a noticeable improvement of 16.6% relative in BLUE score over the phrase-based system, as shown in Table 4.

	BLEU score	Relative Improvement
Internal Test	0.337	16.6%

Table 4: Baseline BLEU scores for hierarchical phrase-based Farsi-to-English SMT system

	BLEU score	Relative Improvement
Internal Test	0.348	3.26%

Table 5: BLEU scores for hierarchical Farsi-to-English SMT system, augmented with Shiraz dictionary

We believe that to a great extent these improvements can be attibuted to the difference in word order between English and Farsi, which are more suitably handled by a hierarchical phrase-based system. For instance, Farsi is an SOV language, which means that the finite verb tends to occur later in the clause than in the corresponding English sentence. This is illustrated in the following sentence pair in which English clause-medial *have* is matched with Farsi clause-final *dArm*:

blh	yh	g#rnAmh	Jdyd	**dArm**
yes	one	passport	new	i.have

'yes I **have** a new passport.'

In the next sections we investigate the result of adding more data resources and applying segmentation into subword units.

3.2. Additional data sources

In addition to the DARPA-supplied Farsi data, we considered the extent to which other freely available online resources could be utilized. To the best of our knowledge, the most extensive previous computational investigation involving Farsi was conducted within the Shiraz project at New Mexico State University ([8]).[2] Among the resources freely available from that project is a bilingual dictionary containing 71,306 Farsi-English entries (52,045 distinct Farsi entries, 23,639 of which being single words). Since the entries are designed to work with the Shiraz morphological analyzer, they are not necessarily in the form most conducive to improving SMT quality. In particular, since the Farsi entries are given in citation form (infinitive for verbs, singular definite for nouns), the entries do not necessarily match the inflected variants found in the training data. As a result, only 9,355 of the Farsi entries are actually found in the training data. Moreover, in terms of adding translations for words not seen in the original training data, the Shiraz dictionary contributes only 123 and 31 new Farsi entries to the coverage of the internal and Eval test sets, respectively.[3] At the same time, it is trivial to add the Farsi-English translation pairs to the training data and subject them to the same preprocessing steps as the original training data.[4] Without any further processing, this leads to a relative improvement of 3.26% in BLEU scores over the original hierarchical phrase-based system, as is shown in Table 5.

	None	PPL=1	PPL=2	PPL=3	PPL=4	PPL=5	PPL=10
Tokens	497558	568393	602736	605113	606429	608903	615047
Types	18502	11928	10935	11010	11094	11189	11313
Suffix types		505	226	123	98	69	18
Prefix types		650	269	133	86	52	42
Segmented word tokens		62297	94115	96356	97776	99387	106537
Segmented word types		8535	9271	9050	8922	8868	8886

Table 6: Data statistics of MORFESSOR-induced segmentations for various perplexity threshold (PPL) settings

	None	PPL=1	2	3	4	5	10
Int. Test	.339	.347	.347	.348	.355	.344	.349

Table 7: BLEU scores for Farsi-to-English SMT system, for various MORFESSOR segmentations

	None	PPL=1	2	3	4	5	10
Int. Test	.348	.356	.357	.360	.362	.355	.359

Table 8: BLEU scores for Farsi-to-English SMT system, for various MORFESSOR segmentations for training data augmented with Shiraz dictionary

3.3. Unsupervised morphological segmentation

One of the challenges for SMT is the fact that languages differ in terms of what information content is packaged into individual words. What gets expressed as a single word in one language may correspond to a series of words in the other. Inasmuch as it is possible to reduce such discrepancies, the quality of word alignments is likely to improve. For instance, Farsi *dvstm* is most naturally translated into English 'my friend'; hence, one way to achieve a closer correspondence between English and Farsi word units is to split the Farsi expression into noun and possessive pronoun segments:

dvst dvst**m** ⇒ dvst **-m**
friend **my** friend

A reliable morphological analyzer for a new language usually cannot be developed without considerable investment of effort and linguistic expertise, which may not be available for a rapid development task. This was in fact the case for Farsi,[5] so we turned to unsupervised methods of detecting subword units. While such methods yield segmentations that often do not correlate very well with more linguistically sound segmentations, they have the obvious benefit of being applicable even if little or nothing is known about the morphology of a new language. Whether automatically derived segmentation leads to any improvement over the segmentation-less baseline system can be determined rather quickly and without having to settle on the type of morphological segmentation to be used (i.e., what kinds of morphemes of what kinds of lexical categories should be separated).

To this end we adopted the MORFESSOR utility ([10]),[6] specifically the MORFESSOR Categories-MAP algorithm ([11]) as a way to obtain a segmentation of the data into morpheme-like units ("morphs"). In particular we used the training set to train a segmentation model that then was applied to the remaining data allowing for the segmentation of unseen words on the basis of the data seen earlier. One of the parameters affecting MORFESSOR segmentation behavior is the *perplexity threshold* (PPL), which, roughly speaking, regulates the aggressiveness with which affixes are postulated. In addition to the default setting of 10, we explored other values and found settings lower than 10 to be more effective for this particular task.[7] Table 3.3 shows the effect of MORFESSOR segmentations for different PPL settings on the training set, with the column headed "None" illustrating the original unsegmented data for comparison. In general, higher PPL values correspond to a higher number of segmented tokens. However, while MORFESSOR segmentation leads to a 35–40% vocabulary reduction over the nonsegmented texts, greater PPL numbers do not necessarily amount to smaller vocabularies. At the same time, greater PPL numbers do result in a smaller affix inventory used in the segmentation.

Table 7 shows the results of various segmentations on BLUE scores. At the MORFESSOR default PPL setting of 10, there is basically no improvements baseline. Better results, however, can be found among the lower PPL settings, in particular at the PPL setting of 4, which yields a relative improvement of 2.0%.

The next experiment involved applying MORFESSOR-derived segmentations on the original training set augmented with the Shiraz dictionary. For systems trained on these sets we observe a improvements over the nonsegmented system for each PPL setting. The best performance of 0.362 can be observed at PPL=4, which constitutes a relative improvement of 2% over the corresponding MORFESSOR-segmented system without the Shiraz dictionary.

The cumulative effect of these measures is a relative improvement over the original standard phrase-based system in BLEU scores of about 25.33%.

3.4. English to Farsi

While most of our effort has been focused on building and improving a system for Farsi-to-English translation, a certain amount of work was also devoted to the other direction. While the absolute scores are considerably lower than for the Farsi-to-English system, we see the same trendlines as before, i.e.,

[2] http://crl.nmsu.edu/Resources/lang_res/persian.html

[3] All but nine of the new entries occur only once in either test set.

[4] This method was inspired by [9], who used a raw bilingual dictionary to improve the performance of their Cebuano system.

[5] Attempts to utilize the Shiraz morphological analyzer remained inconclusive.

[6] www.cis.hut.fi/projects/morpho/morfessorcatmapdownloadform.shtml

[7] Lagus and Creutz suggest that the proper setting is a function of the data size and that larger training corpora require higher settings for optimal performance.

SMT system	Eval test
Standard phrase-based	0.216
Hierarchical phrase-based	0.220
Hierarchical + Shiraz	0.225

Table 9: BLEU scores for English to Farsi SMT systems

improvements for the hierarchical phrase-based system as well for added Shiraz entries, as shown in Table 9. Experiments involving MORFESSOR-induced segmentations for Farsi have so far remained inconclusive.

4. Conclusions

As our results indicate, moving from stardard to hierarchical phrase-based SMT can result in significant performance improvements for language pairs with different word order patterns. Similarly, we have found that additional general-purpose resources such as dictionaries can be helpful even without any additional morphological adjustments. Finally, we have shown that unsupervised methods of morphological segmentation can indeed help improve the performance of an SMT system. This finding contrasts with that of [12], who report no gain for using fully automatically derived segmentations in SMT tasks involving Nordic languages. Whether or not morphological segmentation leads to any gain might of course depend on the particular language pair chosen; at the same time we suspect that the significantly larger data quantities ($\tilde{8}$60,000 aligned sentences reported in [12]) may also result in a diminished impact for morphological segmentation.

In future work we intend to further explore the utility of unsupervised segmentation methods. In particular, we intend to compare our current results with segmentations applied to both source and target sides. In addition, while MORFESSOR utilizes a single parameter to regulate the segmentation of both prefixes and suffixes, we conjecture that a more fine-grained approach that deals with prefixes and suffixes independently might be able to better match the morphological characteristics of a given language.

5. Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Department of Interior–National Business Center (DOI–NBC) under contract number NBCHD040058.

We also thank Mikko Kurimo and the members of the SRI TransTac team for helpful discussion.

6. References

[1] Precoda, K., Zheng, J., Vergyri, D., Franco, H., Richey, C., Kathol, A., and Kajarekar, S., "IraqComm: A Next Generation Translation System," *Interspeech* 2007 (Antwerp, Belgium), August 2007.

[2] Ganjavi, S., Georgiou, P.G., and Narayanan, S., "ASCII based Transcription Systems for Languages with the Arabic Script: The Case of Persian," *ASRU* 2003, pp. 595–600, 2003.

[3] Zheng, J., "SRInterp: SRI's Scalable Multipurpose SMT Engine," SRI Technical Report, June 2008.

[4] Koehn, P., Och, F. J., and Marcu, D, "Statistical Phrase-Based Translation," *HLT-NAACL* 2003, pp. 127–133, 2003.

[5] Chiang, D., "A Hierarchical Phrase-Based Model for Statistical Machine Translation," *ACL*-2005, 2005.

[6] Chiang, D., "Hierarchical Phrase-Based Translation," *Computational Linguistics* 33(2), 2007.

[7] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., "BLEU: A Method for Automatic Evaluation Of Machine Translation" *ACL*-2002, pp. 311–318, 2002.

[8] Amtrup, J. W., Rad, H. M., Megerdoomian, K. and Zajac, R., "Persian-English Machine Translation: An Overview of the Shiraz Project," NMSU, CRL, Memoranda in Computer and Cognitive Science MCCS-00-319, 2000.

[9] Oard, D. W. and Och, F. J., "Rapid-Response Machine Translation for Unexpected Languages," in *Proc. MT Summit IX* (New Orleans, LA), 2003.

[10] Creutz, M. and Lagus, K., "Unsupervised Models for Morpheme Segmentation and Morphology Learning," *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, 2007.

[11] Creutz, M. and Lagus, K., "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text," in Proc. *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (AKRR'05), Espoo, Finland, 2005.

[12] Virpioja, S., Väyrynen, J. J., Creutz, M., and Sadeniemi, M., "Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner," in *Proc. Machine Translation Summit XI*, Copenhagen, Denmark, 10-14 September, 2007, pp. 491-498, 2007.

[13] Goldwater, S. and McClosky, D., "Improving Statistical MT through Morphological Analysis," in *Proc. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, pp. 676–683, 2005.