Unsupervised Morphology Rivals Supervised Morphology for Arabic MT

David Stallard **Jacob Devlin**
Michael Kayser
BBN Technologies
{stallard,jdevlin,rzbib}@bbn.com

Yoong Keok Lee **Regina Barzilay**
CSAIL
Massachusetts Institute of Technology
{yklee,regina}@csail.mit.edu

Abstract

If unsupervised morphological analyzers could approach the effectiveness of supervised ones, they would be a very attractive choice for improving MT performance on low-resource inflected languages. In this paper, we compare performance gains for state-of-the-art supervised vs. unsupervised morphological analyzers, using a state-of-the-art Arabic-to-English MT system. We apply maximum marginal decoding to the unsupervised analyzer, and show that this yields the best published segmentation accuracy for Arabic, while also making segmentation output more stable. Our approach gives an 18% relative BLEU gain for Levantine dialectal Arabic. Furthermore, it gives higher gains for Modern Standard Arabic (MSA), as measured on NIST MT-08, than does MADA (Habash and Rambow, 2005), a leading *supervised* MSA segmenter.

1 Introduction

If unsupervised morphological segmenters could approach the effectiveness of supervised ones, they would be a very attractive choice for improving machine translation (MT) performance in low-resource inflected languages. An example of particular current interest is Arabic, whose various colloquial dialects are sufficiently different from Modern Standard Arabic (MSA) in lexicon, orthography, and morphology, as to be low-resource languages themselves. An additional advantage of Arabic for study is the availability of high-quality supervised segmenters for MSA, such as MADA (Habash and Rambow, 2005), for performance comparison. The MT gain for supervised MSA segmenters on dialect establishes a lower bound, which the unsupervised segmenter must exceed if it is to be useful for dialect. And comparing the gain for supervised and unsupervised segmenters on MSA tells us how useful the unsupervised segmenter is, relative to the ideal case in which a supervised segmenter is available.

In this paper, we show that an unsupervised segmenter can in fact rival or surpass supervised MSA segmenters on MSA itself, while at the same time providing superior performance on dialect. Specifically, we compare the state-of-the-art morphological analyzer of Lee et al. (2011) with two leading supervised analyzers for MSA, MADA and Sakhr[1], each serving as an alternative preprocessor for a state-of-the-art statistical MT system (Shen et al., 2008). We measure MSA performance on NIST MT-08 (NIST, 2010), and dialect performance on a Levantine dialect web corpus (Zbib et al., 2012b).

To improve performance, we apply maximum marginal decoding (Johnson and Goldwater, 2009) (MM) to combine multiple runs of the Lee segmenter, and show that this dramatically reduces the variance and noise in the segmenter output, while yielding an improved segmentation accuracy that exceeds the best published scores for unsupervised segmentation on Arabic Treebank (Naradowsky and Toutanova, 2011). We also show that it yields MT-08 BLEU scores that are higher than those obtained with MADA, a leading *supervised* MSA segmenter. For Levantine, the segmenter increases BLEU score by 18% over the unsegmented baseline.

[1] http://www.sakhr.com/Default.aspx

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 322–327,
Jeju, Republic of Korea, 8-14 July 2012.

2 Related Work

Machine translation systems that process highly inflected languages often incorporate morphological analysis. Some of these approaches rely on morphological analysis for pre- and post-processing, while others modify the core of a translation system to incorporate morphological information (Habash, 2008; Luong et al., 2010; Nakov and Ng, 2011). For instance, factored translation Models (Koehn and Hoang, 2007; Yang and Kirchhoff, 2006; Avramidis and Koehn, 2008) parametrize translation probabilities as factors encoding morphological features.

The approach we have taken in this paper is an instance of a segmented MT model, which divides the input into morphemes and uses the derived morphemes as a unit of translation (Sadat and Habash, 2006; Badr et al., 2008; Clifton and Sarkar, 2011). This is a mainstream architecture that has been shown to be effective when translating from a morphologically rich language.

A number of recent approaches have explored the use of unsupervised morphological analyzers for MT (Virpioja et al., 2007; Creutz and Lagus, 2007; Clifton and Sarkar, 2011; Mermer and Akın, 2010; Mermer and Saraclar, 2011). Virpioja et al. (2007) apply the unsupervised morphological segmenter Morfessor (Creutz and Lagus, 2007), and apply an existing MT system at the level of morphemes. The system does not outperform the word baseline partially due to the insufficient accuracy of the automatic morphological analyzer.

The work of Mermer and Akın (2010) and Mermer and Saraclar (2011) attempts to integrate morphology and MT more closely than we do, by incorporating bilingual alignment probabilities into a Gibbs-sampled version of Morfessor for Turkish-to-English MT. However, the bilingual strategy shows no gain over the monolingual version, and neither version is competitive for MT with a supervised Turkish morphological segmenter (Oflazer, 1993). By contrast, the unsupervised analyzer we report on here yields MSA-to-English MT performance that equals or exceed the performance obtained with a leading supervised MSA segmenter, MADA (Habash and Rambow, 2005).

3 Review of Lee Unsupervised Segmenter

The segmenter of Lee et al. (2011) is a probabilistic model operating at word-type level. It is divided into four sub-model levels. **Model 1** prefers small affix lexicons, and assumes that morphemes are drawn independently. **Model 2** generates a latent POS tag for each word type, conditioning the word's affixes on the tag, thereby encouraging compatible affixes to be generated together. **Model 3** incorporates token-level contextual information, by generating word tokens with a type-level Hidden Markov Model (HMM). Finally, **Model 4** models morphosyntactic agreement with a transition probability distribution, encouraging adjacent tokens with the same endings to also have the same final suffix.

4 Applying Maximum Marginal Decoding to Reduce Variance and Noise

Maximum marginal decoding (Johnson and Goldwater, 2009) (MM) is a technique which assigns to each latent variable the value with the highest marginal probability, thereby maximizing the expected number of correct assignments (Rabiner, 1989). Johnson and Goldwater (2009) extend MM to Gibbs sampling by drawing a set of N independent Gibbs samples, and selecting for each word the most frequent segmentation found in them. They found that MM improved segmentation accuracy over the mean, consistent with its maximization criterion. However, for our setting, we find that MM provides several other crucial advantages as well.

First, MM dramatically reduces the output variance of Gibbs sampling (GS). Table 1 documents the severity of this variance for the MT-08 lexicon, as measured by the average exact-match accuracy and segmentation F-measure between different runs. It shows that on average, 13% of the word tokens, and 25% of the word types, are segmented differently from run to run, which obviously makes the input to MT highly unstable. By contrast the "MM" column of Table 1 shows that two different runs of MM, each derived by combining separate sets of 25 GS runs, agree on the segmentations of over 95% of the word token – a dramatic improvement in stability.

Second, MM reduces noise from the spurious affixes that the unsupervised segmenter induces for large lexicons. As Table 2 shows, the segmenter

Decoding	Level	Rec	Prec	F1	Acc
Gibbs	Type	82.9	83.2	83.1	74.5
	Token	87.5	89.1	88.3	86.7
MM	Type	95.9	95.8	95.9	93.9
	Token	97.3	94.0	95.6	95.1

Table 1: Comparison of agreement in outputs between 25 runs of Gibbs sampling vs. 2 runs of MM on the full MT-08 data set. We give the average segmentation recall, precision, F1-measure, and exact-match accuracy between outputs, at word-type and word-token levels.

	ATB	MT-08		
	GS	GS	MM	Morf
Unique prefixes	17	130	93	287
Unique suffixes	41	261	216	241
Top-95 prefixes	7	7	6	6
Top-95 suffixes	14	26	19	19

Table 2: Affix statistics of unsupervised segmenters. For the ATB lexicon, we show statistics for the Lee segmenter with regular Gibbs sampling (GS). For the MT-08 lexicon, we also show the output of the Lee segmenter with maximum marginal decoding (MM). In addition, we show statistics for Morfessor.

induces 130 prefixes and 261 suffixes for MT-08 (statistics for Morfessor are similar). This phenomenon is fundamental to Bayesian nonparametric models, which expand indefinitely to fit the data they are given (Wasserman, 2006). But MM helps to alleviate it, reducing unique prefixes and suffixes for MT-08 by 28% and 21%, respectively. It also reduces the number of unique prefixes/suffixes which account for 95% of the prefix/suffix tokens (*Top-95*).

Finally, we find that in our setting, MM increases accuracy not just over the mean, but over even the *best*-scoring of the runs. As shown in Table 3, MM increases segmentation F-measure from 86.2% to 88.2%. This exceeds the best published results on ATB (Naradowsky and Toutanova, 2011).

These results suggest that MM may be worth considering for other GS applications, not only for the accuracy improvements pointed out by Johnson and Goldwater (2009), but also for its potential to provide more stable and less noisy results.

Model	Mean	Min	Max	MM
M1	80.1	79.0	81.5	81.8
M2	81.4	80.2	83.0	82.0
M3	81.4	80.1	82.8	83.2
M4	86.2	85.4	87.2	88.2

Table 3: Segmentation F-scores on ATB dataset for Lee segmenter, shown for each Model level M1–M4 on the Arabic segmentation dataset used by (Poon et al., 2009): We give the mean, minimum, and maximum F-scores for 25 independent runs of Gibbs sampling, together with the F-score from running MM over that same set of runs.

5 MT Evaluation

5.1 Experimental Design

MT System. Our experiments were performed using a state-of-the-art, hierarchical string-to-dependency-tree MT system, described in Shen et al. (2008).

Morphological Analyzers. We compare the Lee segmenter with the supervised MSA segmenter MADA, using its "D3" scheme. We also compare with Sakhr, an intensively-engineered, supervised MSA segmenter which applies multiple NLP technologies to the segmentation problem, and which has given the best results for our MT system in previous work (Zbib et al., 2012a). We also compare with Morfessor.

MT experiments. We apply the appropriate segmenter to split words into morphemes, which we then treat as words for alignment and decoding. Following Lee et al. (2011), we segment the test and training sets jointly, estimating separate translation models for each segmenter/dataset combination.

Training and Test Corpora. Our "Full MSA" corpus is the NIST MT-08 Constrained Data Track Arabic training corpus (35M total, 336K unique words); our "Small MSA" corpus is a 1.3M-word subset. Both are tested on the MT-08 evaluation set. For dialect, we use a Levantine dialectal Arabic corpus collected from the web with 1.5M total, 160K unique words and 18K words held-out for test (Zbib et al., 2012b)

Performance Metrics. We evaluate MT with BLEU score. To calculate statistical significance, we use the boot-strap resampling method of Koehn (2004).

5.2 Results and Discussion

Table 4 summarizes the BLEU scores obtained from using various segmenters, for three training/test sets: Full MSA, Small MSA, and Levantine dialect.

As expected, Sakhr gives the best results for MSA. Morfessor underperforms the other segmenters, perhaps because of its lower accuracy on Arabic, as reported by Poon et al. (2009). The Lee segmenter gives the best results for Levantine, inducing valid Levantine affixes (e.g "hAl+" for MSA's "h*A-Al+", English "this-the") and yielding an 18% relative gain over the unsegmented baseline.

What is more surprising is that the Lee segmenter compares favorably with the supervised MSA segmenters on MSA itself. In particular, the Lee segmenter with MM yields higher BLEU scores than does MADA, a leading supervised segmenter, while preserving almost the same performance as GS on dialect. On Small MSA, it recoups 93% of even Sakhr's gain.

By contrast, the Lee segmenter recoups only 79% of Sakhr's gain on Full MSA. This might result from the phenomenon alluded to in Section 4, where additional data sometimes degrades performance for unsupervised analyzers. However, the Lee segmenter's gain on Levantine (18%) is higher than its gain on Small MSA (13%), even though Levantine has more data (1.5M vs. 1.3M words). This might be because dialect, being less standardized, has more orthographic and morphological variability, which unsupervised segmentation helps to resolve.

These experiments also show that while Model 4 gives the best F-score, Model 3 gives the best MT scores. Comparison of Model 3 and 4 segmentations shows that Model 4 induces a much larger number of inflectional suffixes, especially the feminine singular suffix "-p", which accounts for a plurality (16%) of the differences by token. While such suffixes improve F-measure on the segmentation references, they do not correspond to any English lexical unit, and thus do not help alignment.

An interesting question is how much performance might be gained from a supervised segmenter that was as intensively engineered for dialect as Sakhr was for MSA. Assuming a gain ratio of 0.93, similar to Small MSA, the estimated BLEU score would be 20.38, for a relative gain of just 5% over the unsupervised segmenter. Given the large engineering effort that would be required to achieve this gain, the unsupervised segmenter may be a more cost-effective choice for dialectal Arabic.

System		Small MSA	Full MSA	Lev Dial
Unsegmented		38.69	43.45	17.10
Sakhr		***43.99***	***46.51***	19.60
MADA		43.23	45.64	19.29
Morfessor		42.07	44.71	18.38
Lee GS	M1	43.12	44.80	19.70
	M2	43.16	45.45	***20.15+***
	M3	43.07	44.82	19.97
	M4	42.93	45.06	19.55
Lee MM	M1	43.53	45.14	19.75
	M2	43.45	45.29	19.75
	M3	**43.64+**	**45.84**	**20.09**
	M4	43.56	45.16	19.93

Table 4: BLEU scores for all experiments. Full MSA is the the full MT-08 corpus, Small MSA is a 1.3M-word subset, Lev Dial our Levantine dataset. For each of these, the highest Lee segmenter score is in bold, with "+" if statistically significant vs. MADA at the 95% confidence level or higher. The highest overall score is in bold italic.

6 Conclusion

We compare unsupervised vs. supervised morphological segmentation for Arabic-to-English machine translation. We add maximum marginal decoding to the unsupervised segmenter, and show that it surpasses the state-of-the-art segmentation performance, purges the segmenter of noise and variability, yields BLEU scores on MSA competitive with those from supervised segmenters, and gives an 18% relative BLEU gain on Levantine dialectal Arabic.

Acknowledgements
This material is based upon work supported by DARPA under Contract Nos. HR0011-12-C00014 and HR0011-12-C00015, and by ONR MURI Contract No. W911NF-10-1-0533. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US government. We thank Rabih Zbib for his help with interpreting Levantine Arabic segmentation output.

References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*.

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4:3:1–3:34, February.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of ACL*.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparametric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of EMNLP-CoNLL*, pages 868–876.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2011. Modeling syntactic context improves morphological segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Coşkun Mermer and Ahmet Afşın Akın. 2010. Unsupervised search for the optimal segmentation for statistical machine translation. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 31–36, Uppsala, Sweden, July. Association for Computational Linguistics.

Coşkun Mermer and Murat Saraclar. 2011. Unsupervised Turkish morphological segmentation for statistical machine translation. In *Workshop on Machine Translation and Morphologically-rich languages*, January.

Preslav Nakov and Hwee Tou Ng. 2011. Translating from morphologically complex languages: A paraphrase-based approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Jason Naradowsky and Kristina Toutanova. 2011. Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

NIST. 2010. NIST 2008 Open Machine Translation (Open MT) Evaluation. `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T21/`.

Kemal Oflazer. 1993. Two-level description of Turkish morphology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*.

Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of the Machine Translation Summit XI*.

Larry Wasserman. 2006. *All of Nonparametric Statistics*. Springer.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of EACL*.

Rabih Zbib, Michael Kayser, Spyros Matsoukas, John Makhoul, Hazem Nader, Hamdy Soliman, and Rami Safadi. 2012a. Methods for integrating rule-based and statistical systems for Arabic to English machine translation. *Machine Translation*, 26(1-2):67–83.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012b. Machine translation of Arabic dialects. In *NAACL 2012: Proceedings of the 2012 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, Quebec, Canada, June. Association for Computational Linguistics.