A Little Linguistics Goes a Long Way: Unsupervised Segmentation with Limited Language Specific Guidance

Alexander Erdmann, Salam Khalifa, Mai Oudah, Nizar Habash and Houda Bouamor[†]
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
[†]Carnegie Mellon University in Qatar, Qatar
{ae1541,salamkhalifa,mai.oudah,nizar.habash}@nyu.edu
hbouamor@cmu.edu

Abstract

We present de-lexical segmentation, a linguistically motivated alternative to greedy or other unsupervised methods, requiring language specific knowledge, but no direct supervision. Our technique involves creating a small grammar of closed-class affixes which can be written in a few hours. The grammar over generates analyses for word forms attested in a raw corpus which are disambiguated based on features of the linguistic base proposed for each form. Extending the grammar to cover orthographic, morpho-syntactic or lexical variation is simple, making it an ideal solution for challenging corpora with noisy, dialect-inconsistent, or otherwise non-standard content. We demonstrate the utility of de-lexical segmentation on several dialects of Arabic. We consistently outperform competitive unsupervised baselines and approach the performance of state-of-the-art supervised models trained on large amounts of data, providing evidence for the value of linguistic input during preprocessing.

1 Introduction

Non-standard domains, dialectal variation, and unstandardized spelling make segmentation challenging, though morphologically rich languages require good segmentation to enable downstream applications from syntactic parsing to machine translation (MT). For domains lacking sufficient annotated data to train segmenters, one must resort to language specific greedy techniques or language agnostic unsupervised techniques. Greedy techniques use maximum matching to identify base words, leveraging large dictionaries (Guo, 1997). Yet such dictionaries are often unavailable or too expensive for low resource languages. Language agnostic unsupervised options like MORFESSOR (Creutz and Lagus, 2005) and byte pair encoding (BPE) (Sennrich et al., 2016) assume no resources beyond raw text but can yield lower performance on downstream tasks (Vania and Lopez, 2017; Kann et al., 2018). They also suffer from typological biases and favor intended applications at the expense of others.

To this end, we present *De-lexical Segmentation* (DESEG), a slightly more expensive but powerful alternative to language agnostic morphological segmentation, realizing most of the benefits of supervised segmentation at far less a cost. DESEG requires language specific input in the form of a small grammar describing the combinatorics of closed-class affixes. We demonstrate that such a grammar can be constructed easily and rapidly for a new language or dialect. Hence, DESEG addresses the scenario in which there is no supervised segmenter available for a given language or dialect (or no segmenter trained on a domain with sufficient lexical overlap with the target domain in its training data), but the user does have linguistic knowledge of the target language/dialect.

The user-provided grammar is employed in conjunction with a large, raw corpus. The grammar over generates analyses for all words therein, allowing for maximal recall not only of the possible affix combinations, but also variant spellings and dialectal idiosyncrasies. The preferred analysis is disambiguated based on the fertility with which its proposed base attaches to different affixes in analyses of other words throughout the corpus. This follows from the logic that valid bases are more likely to productively combine with more exponents[1] (Bertram et al., 2000). By leveraging language specific resources but learning to disambiguate empirically without supervision, we mitigate much of the sparsity inherent in processing

[1] Exponents refer to recurring means by which morpho-syntactic properties are realized within classes of words, e.g., adding suffix +*s* to get the third person singular present tense for verbs like WALK, TALK, and SKIP.

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 113–124
Florence, Italy. August 2, 2019 ©2019 Association for Computational Linguistics

non-standard domains.

Using a corpus of several Arabic dialects exhibiting rich and complex morphology, unstandardized spelling, and variation bordering on mutual unintelligibility, we evaluate DESEG intrinsically on language modeling (LM) and extrinsically on MT. DESEG consistently outperforms MORFESSOR and BPE while only costing a few hours of grammar-building labor; and in some environments it outperforms state-of-the-art supervised Arabic tokenizers MADAMIRA (Pasha et al., 2014) and FARASA (Abdelali et al., 2016). The success of such a simple model is strong evidence for the value of linguistic input during preprocessing. DESEG is publicly available at `github.com/CAMeL-Lab/deSeg`.

2 Related Work

Many morphologically rich languages lack crucial preprocessing resources like morphological analyzers or segmenters. Even well resourced languages often lack such resources for non-standard dialects and domains. There have been many approaches to address this problem, varying along a number of dimensions: the degree of language independence or specificity, the required amount of machine learning supervision, the degree of depth and richness of the morphological representations.

Language agnostic unsupervised models There are many works using minimally supervised to unsupervised models of morphology for connecting morphologically related words and identifying optimal (and at times application dependent) segmentations (Smith and Eisner, 2005; Creutz and Lagus, 2005; Snyder and Barzilay, 2008; Poon et al., 2009; Dreyer and Eisner, 2011; Stallard et al., 2012; Sirts and Goldwater, 2013; Narasimhan et al., 2015; Sennrich et al., 2016; Eskander et al., 2016b; Ataman et al., 2017; Ataman and Federico, 2018; Eskander et al., 2018). In this paper, we compare to two popular language agnostic segmentation systems: MORFESSOR (Creutz and Lagus, 2005) and BPE (Sennrich et al., 2016). Both train on large corpora of unannotated text in an unsupervised manner.

Standard Arabic models Modern Standard Arabic (MSA) morphological analysis, disambiguation and tokenization has been the focus of a large number of efforts. Khoja and Garside (1999) was one of the earliest published efforts on automatic shallow and deterministic segmentation for MSA. Darwish (2002) used limited resources and greedy techniques to automatically learn rules and statistics to build a shallow morphological analyzer. There are many MSA morphological analyzers with rich representations and good coverage that required very intensive efforts to create (Beesley, 1998; Buckwalter, 2004; Attia, 2006, 2007; Smrž, 2007; Boudchiche et al., 2017). Buckwalter (2004) is perhaps the most commonly used among them, as it contributed the representations for the Penn Arabic treebank (PATB) (Maamouri and Bies, 2004). The PATB has been the most used resource for supervised morphological disambiguation (Diab et al., 2004; Habash and Rambow, 2005; Pasha et al., 2014; AlGahtani and McNaught, 2015; Zalmout and Habash, 2017). Some efforts have used other annotated resources and/or large unannotated data sets (Lee et al., 2003; Abdelali et al., 2016; Freihat et al., 2018). More closely related to this paper, Erdmann and Habash (2018) demonstrated that de-lexicalized information provides a cheap means of inducing morphological knowledge and thereby predicting lexical information in MSA. They employ a de-lexicalized grammar which is similar to ours, but they do not handle dialectal variants or spelling variation. They also do not use the grammar for segmentation, but for pruning word embedding clusters in order to predict the paradigm membership of forms encountered in raw text.

Dialectal Arabic models Work on dialectal Arabic morphology and tokenization is relatively newer than work on MSA. Some of the earlier efforts worked on rule-based approaches to model dialectal morphology directly (Habash and Rambow, 2006; Habash et al., 2012), or exploiting existing MSA resources (Salloum and Habash, 2014). Later, a number of annotation efforts have led to the creation of varying sizes of dialectal annotated corpora following the style of the PATB (Maamouri et al., 2014; Jarrar et al., 2016; Al-Shargi et al., 2016; Khalifa et al., 2018; Alshargi et al., 2019). The created annotations supported models for dialectal Arabic analysis, disambiguation and tokenization building on the same successful approaches in MSA (Eskander et al., 2016a; Habash et al., 2013; Pasha et al., 2014; Zalmout et al., 2018; Zalmout and Habash, 2019). More closely related to this paper, Eldesouki et al. (2017) used de-lexicalized analysis strategy for four colloquial varieties of Arabic, though they also use minimal training data and extract features from an open class lexicon to learn either an SVM or bi-LSTM-CRF disambiguation model. They further show that domain adaptation from existing MSA training data is beneficial. Also, Samih et al. (2017) applied a related model to segmentation, allowing different Arabic dialects to inform one another, thus avoiding the need to perform dialect identification during preprocessing.

We compare our model to MADAMIRA (Pasha et al., 2014) and FARASA (Abdelali et al., 2016), which represent the fully supervised state of the art for segmenting Arabic in the standard domain, but have limited support for multiple colloquial variants of the language.

Finally, we note that, linguistically, our work is inspired by Bertram et al. (2000) who find that *prolific* stems with large derivational families are accessed more quickly. Their work suggests that stem *fertility*—or the productivity with which a stem can combine with different affixes—is cognitively relevant to morphological organization.

3 De-lexical Segmentation for Arabic

In this section, we introduce a case study on segmenting a multi-dialect Arabic corpus and explain the linguistic challenges it presents for popular approaches to segmentation. Furthermore, we discuss the construction of DESEG's grammar and its disambiguation algorithm.

3.1 Arabic and its Dialects

Arabic is highly diaglossic (Ferguson, 1959), with the relatively consistent high register of Modern Standard Arabic being learned in schools across the Arab World. Meanwhile the often mutually unintelligible low register variants—collectively known as dialectal Arabic (DA)—are spoken colloquially. The phonological, morpho-syntactic, and lexical variation within the Arabic sprachbund is comparable to that among Romance languages (Chiang et al., 2006; Rouchdy, 2013; Erdmann et al., 2017), leading to problematic noise in multi-dialect corpora (Erdmann et al., 2018). Furthermore, lack of spelling conventions in DA exacerbates data sparsity, as does a rich morphology featuring templatic phenomena and robust cliticization, making it challenging to train quality segmenters even with much supervised data.

3.2 Data

To demonstrate how our model handles such challenging phenomena, we apply it to the CORPUS6 subset of the MADAR-BTEC (Takezawa et al., 2002) corpus of Arabic dialects (Salameh et al., 2018). This consists of 12,000 sentences in the travel domain (9,000 for training) parallel between English, MSA, and the DA varieties spoken in Beirut, Cairo, Doha, Rabat, and Tunis. This comprises a representative sample of the breadth of intra-DA variation (Bouamor et al., 2018).

In addition to CORPUS6, we also use large amounts of raw monolingual data to train our segmenter and the unsupervised baselines. To avoid introducing even more noise, we restrict our monolingual datasets as much as possible to similar domains. For DA, we use the four subsets of Almeman and Lee (2013)'s web crawl of forums, comments and blogs, consisting of over 10 million words for each subset's dialect region. It is worth noting however, that the granularity of their dialect regions is coarser than the granularity of CORPUS6. Hence, their Maghrebi dialect corresponds to two dialects in CORPUS6, Tunis and Rabat, while the remaining three dialect regions have rather obvious one-to-one correspondences with CORPUS6, i.e., Egyptian to Cairo, Levantine to Beirut, and Gulf to Doha. For MSA, which rarely occurs consistently (i.e., outside of brief instances of code-mixing) in such casual domains, we used the TED corpus (Cettolo and Girardi, 2012) for our monolingual data set, finding a compromise between domain relevance and corpus size. It contains about 2.5 million words.

Obviously, CORPUS6 is small relative to other MT corpora, but this is exactly why it is a meaningful evaluation corpus. Larger parallel corpora are often only available for better resourced languages/domains where fully supervised segmenters are also more likely to be available, negating the need to build one's own segmenter. Furthermore, as parallel data becomes less sparse, tokenization necessarily has less of an effect since models can memorize and effectively use longer sequences. With that said, CORPUS6 is commissioned, and in future work we would like to also test DESEG's performance on natural corpora.

3.3 De-lexical Analysis

The DESEG grammar provides all possible *de-lexical* analyses of words by assuming any n-gram

(A)

Morph.Feat.	Prefix	Suffix
PV.1US	∅	+t +ت
PV.1UP	∅	+nA +نا
PV.2MS	∅	+t +ت
PV.2FS	∅	+t/ty +ت/تي
PV.2US	∅	+ty +تي
PV.2UP	∅	+twA +توا
PV.3MS	∅	∅
PV.3FS	∅	+t +ت
PV.3UP	∅	+wA +وا
IV.1US	A/n+ ا/ن+	∅
IV.1UP	n+ ن+	+wA/∅ +وا/∅
IV.2MS	t+ ت+	∅
IV.2FS	t+ ت+	+y/yn/∅ +ي/ين/∅
IV.2UP	t+ ت+	+wA/wn +وا/ون
IV.3MS	y+ ي+	∅
IV.3FS	t+ ت+	∅
IV.3UP	y+ ي+	+wA/wn +وا/ون
CV.2MS	∅	∅
CV.2FS	∅	+y/∅ +ي/∅
CV.2UP	∅	+wA +وا
NOM.MS	∅	∅
NOM.FS	∅	+$\hbar$ +ة
NOM.MD	∅	+yn +ين
NOM.FD	∅	+tyn +تين
NOM.MP	∅	+yn +ين
NOM.FP	∅	+At +ات
PART	∅	∅

(B)

Proclitics	Orth	POS
ART	Al+ ال+	DET
ART	h+Al+ ه+ال+	DEM_PART+DET
PART$_n$	š+ ش+	INTERROG_PART
PART$_n$	ς+ ع+	PREP
PART$_n$	b+ ب+	PREP
PART$_n$	d+ د+	PREP
PART$_n$	f+ ف+	PREP
PART$_n$	k+ ك+	PREP
PART$_n$	w+ و+	PREP
PART$_n$	yA+ يا+	VOC_PART
PART$_n$	Ā/A+ آ/ا+	VOC_PART
PART$_n$	l+ ل+	PREP
PART$_v$	H+ ح+	FUT_PART
PART$_v$	b/m+ ب/م+	PROG_PART
PART$_v$	b+ ب+	FUT_PART
PART$_v$	g+ غ+	FUT_PART
PART$_v$	h+ ه+	FUT_PART
PART$_v$	k+ ك+	PROG_PART
PART$_v$	t+ ت+	PROG_PART
PART$_v$	l+ ل+	JUS_PART
m_NEG	m/mA+ م/ما+	NEG_PART
CONJ	f+ ف+	CONJ
CONJ	f+ ف+	CONNEC_PART
CONJ	f+ ف+	RC_PART
CONJ	t+ ت+	SUB_CONJ
CONJ	w+ و+	CONJ
CONJ	w+ و+	SUB_CONJ

(C)

Enclitics	Orth	POS
PRON$_{n,v}$	+kw +كو	2UP
PRON$_{n,v}$	+ky +كي	2UP
PRON$_{n,v}$	+km +كم	2MP/2UP
PRON$_{n,v}$	+h/w +ه/و	3MS
PRON$_{n,v}$	+hA +ها	3FS
PRON$_{n,v}$	+hm +هم	3UP
PRON$_{n,v}$	+hn +هن	3FP
PRON$_{n,v}$	+hn/n +ن/هن	3UP
PRON$_{n,v}$	+j +ج	2FS
PRON$_{n,v}$	+k +ك	2MS/2FS
PRON$_{n,v}$	+kn +كن	2UP/2FP
PRON$_{n,v}$	+nA +نا	1UP
PRON$_{n,v}$	+y +ي	1US
PRON$_v$	+ny +ني	1US
IOBJ	+l+h/w +ل+ه/و	PREP+3MS
IOBJ	+l+hA +ل+ها	PREP+3FS
IOBJ	+l+hm +ل+هم	PREP+3MP/3UP
IOBJ	+l+hn/n +ل+هن/ن	PREP+3FP/3UP
IOBJ	+l+j +ل+ج	PREP+2FS
IOBJ	+l+k +ل+ك	PREP+2MS/2FS
IOBJ	+l+km +ل+كم	PREP+2MP/2UP
IOBJ	+l+kn +ل+كن	PREP+2FP/2UP
IOBJ	+l+nA +ل+نا	PREP+1UP
IOBJ	+l+y +ل+ي	PREP+1US
NEG_PART	+š +ش	NEG_PART

$$
\begin{array}{rrcl}
\textbf{(D)} & \textbf{WORD} & \rightarrow & \text{CONJ? } (\textbf{NOM}|\textbf{VERB}|\textbf{PART}) \\
 & \textbf{PART} & \rightarrow & \textbf{PART}_0 \ \text{PRON}_n\text{?} \\
 & \textbf{NOM} & \rightarrow & \text{PART}_n\text{? (ART? } \textbf{NOM}_0|\textbf{NOM}_0 \ \text{PRON}_n\text{?)} \\
 & \textbf{VERB} & \rightarrow & \text{m\_NEG? } \textbf{VERB}_1 \ \text{NEG\_PART?} \\
 & \textbf{VERB}_1 & \rightarrow & (\text{PART}_v\text{? } \textbf{VERB}_{0.iv}|\textbf{VERB}_{0.pv}|\textbf{VERB}_{0.cv}) \ \text{PRON}_v\text{? IOBJ?}
\end{array}
$$

Table 1: All the elements needed to build a de-lexicalized morphological analyzer for the five dialects. (A) represents all the abstract meta paradigms for the basic Arabic POS: verbs (perfective (PV), imperfective (IV), and command (CV)), nominals (NOM), and particles (PART). (B) and (C) are the set of clitics along with their respective POS, categorized by their morphological role. The CFG in (D) describes the valencies of the clitics surrounding the base form.

of some minimum length can be an open class base, provided the remaining characters comprise a supported affix pattern. Hence, a simple grammar which only supports words without affixes or with a single suffix, +*s*, would return two analyses for *wugs*: *wugs* and *wug* +*s*, and one for *foo*: *foo*. To build such a grammar for an Arabic dialect, we target clitic affixation, as this phenomenon is non-templatic with minimal fusional edits, making it easier to model with a smaller grammar, yet it accounts for a great deal of sparsity, as Arabic clitics are as productive as regular inflectional exponents.

We use our grammar to build a de-lexicalized morphological analyzer for all DA dialects targeting the D3 segmentation scheme (Habash, 2010), which separates all clitics and only clitics from the base forms to which they attach. We chose D3 as Sadat and Habash (2006) demonstrate it to be the most effective scheme for low resource Arabic MT. [2] While Arabic exhibits many other non-concatenative, templatic phenomena which complicate segmentation and tokenization, clitics are always concatenated to the outsides of base forms after the templatic pattern has been applied and are thus easier to separate. Occasionally, fusional processes can alter phonemes/graphemes on either side of base–clitic or clitic–clitic boundaries, but no templatic process is ever invoked to alter the internal structure of bases by affixing any clitic.

We follow Khalifa et al. (2017)'s approach to

[2] With more data, the more effective schemes are ATB and D2 (Sadat and Habash, 2006). ATB resembles D3 but does not separate the definite article proclitic. D2 resembles ATB but does not separate the pronominal enclitic.

extending paradigms with possible clitic combinations, though we don't require any stem lexical information. Hence, we cheaply enable the grammar to over generate, accommodating more spelling variants and removing the need to construct an open class lexicon. Instead, we simply provide meta paradigms for abstractions over base forms with the same combinatorics. Each cell in a meta paradigm represents a unique exponent, or possible mapping of clitics to positions surrounding the abstract base, such that the inflected form would be valid for any real base represented by that meta paradigm. Considering verbal affixation in English, *walk* and *talk* would be two real bases taking the same meta paradigm with four cells, represented by exponents _+*ing* _+*s*, _, and _+*ed*. Thus, any two bases exhibiting distinct exponent signatures will belong to distinct meta paradigms.

In Arabic, by contrast, paradigms are enumeratively and integratively more complex than the TALK/WALK meta paradigm (Ackerman and Malouf, 2013). Table 1[3] exemplifies Arabic's enumerative complexity, as verbs, for instance, depending on dialect, can take some 20 affixes according to (A), realizing various combinations of aspect, person, gender, and number.

Having taken an affix, the verb can participate in myriad possible additional combinations with clitics in (B) and (C) as dictated by the bottom two rules in the CFG in (D). Arabic is thus, integratively complex in that rich exponents can be comprised of many interacting morphemes whose meanings are often affected by each other's presence. Furthermore, fusional processes acting on such complex forms results in frequent allomorphy. Allomorphy is mostly limited to internal, non-clitic morphemes, which enables us to greatly reduce sparcity without propagating error by focusing on clitics. Hence, we can represent all verbs with a single meta paradigm which is large, but can be described in two CFG rules. In practice then, each of the 20 possible affixes in (A) will correspond to distinct abstract bases, though this eliminates the need to specify 20 distinct meta paradigms for single lexemes. We target relating these abstract bases to each other via non-concatenative modeling in future work.

In terms of the effort required to create the grammar, there are a total of 98 unique affixes for all dialects. We include the non-clitic affixes in Table 1 (A) in this count as they are used to restrict the set of possible meta paradigms. Of these, 45% appear in at least two dialects and 33% appear in all dialects. The total number of affix–dialect pairs is 288. On average, 88% of each dialect's affixes are shared by at least one other dialect and 45% by all dialects. The average dialect specific list contains 58 affixes and adding a second dialect requires an additional 16. Adding a third, fourth, and fifth dialect requires 10, 8, and 7 additional affixes on average, respectively. Thus, building a single dialect grammar is cheap and adding dialects is even cheaper. Our final grammar contains five meta paradigms, one for each of the basic Arabic parts-of-speech—verbs (PV, IV, and CV), nominals, and particles—compiled into an analyzer like that of Buckwalter (2004).

3.4 Unsupervised Disambiguation

DESEG supports two simple, fast models for disambiguating the grammar's analyses. The first, DESEG$_g$, greedily selects the maximum match analysis, or that with the smallest base after matching affixes. The second, DESEG$_f$, selects the analysis with the most fertile base. The fertility of each candidate base is calculated in the raw corpus by counting the possible combinations of adjacent affixes with which it appears over all analyses for all words in which it is proposed as a base.

For example, consider the three-word toy corpus in Table 2. بيقولها *byqwlhA*, correctly segmented as *b*+ *yqwl* +*hA*, PROG+ say.3MS +it, 'he is saying it', has six possible analyses, each with a different candidate base. Two candidate bases, *yqwl* and *byqwl*, are also candidate bases for another word, بيقول *byqwl* 'he's not saying', but only *yqwl* exhibits multiple unique adjoining affix sets. In *byqwlhA*, it takes the circumfix *b* | *hA*, while in *byqwl*, it takes the prefix *b*. The fertility of base *yqwl* suggests it is more likely to be a productive stem in the language, whereas the lack of fertility for the base *byqwl* suggests it is not systematically utilized in the language as a base might be expected to be used, and that it is more likely a simple coincidence that enables the over permissive grammar to allow such a candidate.

The final word in the vocabulary, يبقولي *ybqwly*, correctly segmented as *ybqw* +*l* +*y*, remain.3MP +to +me 'they remain for me', is challenging because no other inflection of the lexeme is attested.

[3] POS tags in Table 1 are presented in the Buckwalter scheme used in annotating the Penn Arabic Treebank (PATB) (Maamouri and Bies, 2004)

Vocabulary	Candidate Segmentations	Candidate Bases	Attested Adjoining Affixes	Fertility	Base Length
byqwlhA بيقولها	***b+ yqwl +hA*** **ب+ يقول +ها**	***yqwl*** **يقول**	***b* | ∅ , *b* | *hA*** **ب | ∅, ب | ها**	**2**	**4**
	byqwlhA بيقولها	*byqwlhA* بيقولها		0	7
	b+ yqwlhA ب+ يقولها	*yqwlhA* يقولها	*b* | ∅ ب | ∅	1	6
	b+ yqw +l +hA ب+ يقو +ل +ها	*yqw* يقو	*b* | *l* ب | ل	1	3
	byqwl +hA بيقول +ها	*byqwl* بيقول	∅ | *hA* ها | ∅	1	5
	byqw +l +hA بيقو +ل +ها	*byqw* بيقو	∅ | *l* ل | ∅	1	4
byqwl بيقول	***b+ yqwl*** **ب+ يقول**	***yqwl*** **يقول**	***b* | ∅ , *b* | *hA*** **ب | ∅, ب | ها**	**2**	**4**
	byqwl بيقول	*byqwl* بيقول	∅ | *hA* ها | ∅	1	5
ybqwly يبقولي	***ybqw +l +y*** **يبقو +ل +ي**	***ybqw*** **يبقو**	**∅ | *l*** **ل | ∅**	**1**	**4**
	ybqwly يبقولي	*ybqwly* يبقولي		0	6
	ybqwl +y يبقول +ي	*ybqwl* يبقول	∅ | *y* ي | ∅	1	5

Table 2: Calculating fertility in a toy Arabic corpus of three words given all possible candidate analyses of the input corpus vocabulary. Correct analyses are depicted in bold.

Yet, by maximum matching on the affixes, we choose the correct analysis—*ybqw* plus the complex suffix of prepositional *l* followed by object *y*—as the proposed base *ybqw* is shorter than the other candidate base which is produced by erroneously assuming a nominal meta paradigm. The nominal analysis re-analyzes *y* as the first person possessive enclitic and crucially extends the base with *l*, as *l* is not a viable nominal enclitic. Thus, choosing the shortest base can help to eliminate coincidentally feasible analyses.

Each model, DESEG$_f$ and DESEG$_g$, breaks ties using the other. Thus, DESEG$_f$ would correctly segment the entire toy corpus, as the correct analyses in *byqwlhA* and *byqwl* feature the uniquely most fertile candidate bases, and while there is a fertility tie for *ybqwly*, backing off to the candidate segmentation with the smallest base length correctly selects the segmentation with *ybqw* as the base. DESEG$_g$ correctly segments *byqwl* and *ybqwly*, but incorrectly predicts that the stem-final *l* in *byqwlhA* is actually the same enclitic preposition present in *ybqwly* and thus, over segments.

In the event of ties after considering both fertility and base length, both models back off again to the analysis with the base that most frequently occurs as a full word in the raw corpus. Prioritizing this frequency above either fertility or base length minimization always hurt performance, even though it proved quite useful as a feature for Narasimhan et al. (2015). We attribute this seeming discrepancy to the interaction of Arabic's rich morphology with the noise of unstandardized DA data. Many gold bases actually cannot appear as stand-alone words due to the fusional morphology and various writing conventions greatly affect the frequency with which bases that *can* manifest as stand-alone words *actually do*.

4 Evaluation

We compare DESEG to several alternative segmentation models. We use the CORPUS6 dev set to pick the optimal minimum base length on an intrinsic LM perplexity evaluation, and then perform an extrinsic MT evaluation on the test set.

4.1 Models

We evaluate the following models:

PLAIN This baseline segments only punctuation.

MADAMIRA Egyptian and MSA versions are available for MADAMIRA, which disambiguates a rule-based morphological analyzer's output with an SVM trained on morphologically annotated data. We use the Egyptian version as it is pretrained on a superset of the MSA data to capture code switching. Thus, performance does not significantly drop when testing on MSA, and performance is significantly greater when testing on DA varietes—even those far outside of Egypt—due to many shared intra-DA linguistic traits not present in MSA (Khalifa et al., 2017). MADAMIRA is a tokenizer in that it not only segments but also mitigates data sparsity due to allomorphy by recovering the canonical underlying morpheme for each segment. We run MADAMIRA in D3 tokenization mode, facilitating comparison with DESEG.

FARASA Similar to MADAMIRA, FARASA is a pre-trained, SVM-based system leveraging gold annotations and external dictionaries. Together, FARASA and MADAMIRA represent the state of

	Invariable			Trainable					
	Rule-based	Pre-trained		Unsupervised		Unsupervised + De-lexical Grammar			
	PLAIN	MADAMIRA	FARASA	BPE	MORFESSOR	DESEG$_{g3}$	DESEG$_{f3}$	DESEG$_{g2}$	DESEG$_{f2}$
Tokens	42,125	54,559	58,728	53,617	53,509	62823	64708	72644	70704
OOV%	6.8%	3.0%	2.60%	0.7%	2.6%	2.0%	2.1%	1.8%	**1.7%**
Perplexity	163.0	75.0	59	132.2	96.5	52.6	48.0	**33.5**	36.2

Table 3: Out of vocabulary (OOV) and perplexity for all tokenization models in the pooled dialects environment.

Dialects used to Train Segmenter(s)	**Dialects used to Train MT System(s)**	Invariable			Trainable			
		Rule-based	Pre-trained		Unsupervised		Unsupervised + De-lexical Grammar	
		PLAIN	MADAMIRA	FARASA	BPE	MORFESSOR	DESEG$_{g2}$	DESEG$_{f2}$
Pooled	**Pooled**	29.8	31.5	**32.7**	29.9	30.6	32.0	32.3
Individual	**Individual**	28.7	**31.4**	31.2	28.4	30.1	30.9	31.3
Individual	**Pooled**	29.8	31.5	32.7	30.6	31.8	32.5	**32.9**

Table 4: Macro BLEU scores for each tokenization model on CORPUS6 in three environments distinguishing how dialects are pooled or treated separately when training the tokenizer and MT system.

the art for a number of morphological tasks in Arabic. FARASA differs from MADAMIRA in that only one version is publicly available, it segments only, not attempting to tokenize, and the segmentation scheme is linguistically ad hoc, tending to be slightly more granular than D3.

BPE Byte pair encoding uses an algorithm originally designed for file compression to perform unsupervised segmentation. BPE was originally proposed to reduce vocabulary size to make neural MT tractable (Sennrich et al., 2016), as the algorithm's simplicity enables easy application to any language. It separates all characters in the corpus, then performs a pre-determined number of join operations, merging all instances of specified bigrams. Joins are determined such that the resulting corpus will contain as few tokens as possible given the number of join operations allowed. Thus, while the algorithm is unsupervised and easy to apply to any language, it is linguistically naive, assuming that morphological organization is driven solely by enumerative efficiency concerns. Likely for this reason, BPE has not been demonstrated to be particularly useful for applications beyond neural MT (Kann et al., 2018).

MORFESSOR The de facto publicly available unsupervised segmentation system is MORFESSOR. Like BPE, MORFESSOR trains in an unsupervised fashion on large amounts of data and is easily run on any language. Efficient encoding of morphology is also at the center of MORFESSOR's objective function, though it considers not only how compact the corpus can be represented, but also how compact the grammar describing morpheme combinatorics can be represented. Stem morphemes are distinguished from affixal morphemes as the model seeks to limit the number of unique signatures—the sets of unique affixes which can occur with a given stem—that result from the learned segmentation scheme. While MORFESSOR performs well on a number of unsupervised segmentation tasks, it is known to have typological biases toward the languages for which it was originally developed (Kirschenbaum, 2015).

DESEG Our model, described in Section 3, finds a compromise between the convenience of language agnostic unsupervised systems and the performance of systems leveraging language specific resources. DESEG can be run with a minimum base length of either 2 or 3 characters and a priority of base fertility maximization (f) over greedy base length minimization, or vice versa (g). Minimum base length and priority are represented as subscripts in all relevant tables.

4.2 Intrinsic Language Modeling Evaluation

Table 3 shows the LM results for tokenizing CORPUS6 where all trainable segmenters are trained on all of the raw data pooled together instead of training dialect specific tokenizers on relevant subsections of the 40+ million word corpus. To enable pooled DESEG grammars, each dialect's grammar is merged into one highly permissive, over generating pan-Arabic grammar. In the unpooled training scenario, perplexity rankings were consistent with those displayed here. Our model greatly reduces both perplexity and out of vocabulary over all competitive models, though we also exhibit a tendency to over segment. Our best DESEG variants use a minimum base length of two, which is logical because while Arabic features mainly triradical roots, gemination causes many base forms to reduce to only two graphemes. In the intrinsic evaluation, it is difficult to tell whether the preference for greed (DESEG$_{g2}$) or fertility (DESEG$_{f2}$) is better. Our success is likely due to the fact that we alone cover all the dialects, yet that coverage was achieved in a fraction of the time spent constructing the annotated data upon which state-of-the-art systems rely to cover just a single dialect.

4.3 Extrinsic Machine Translation Evaluation

We conduct MT experiments translating Arabic dialects to English in three environments. Pooled–pooled trains segmenters (only trainable segmenters) on the monolingual corpus with all dialects pooled and the MT system on all the dialects pooled. Individual–individual trains six segmenters on relevant subsections of the monolingual data and six MT systems on the relevant partitions of CORPUS6. Individual–pooled trains individual segmenters but one pan-Arabic MT system, which is reasonable to reduce the over generation of the morphological model but leverage shared information during MT. Neural MT has been used with dialects (Hassan et al., 2017), but given the extreme scarcity of in-domain data, statistical MT (Koehn et al., 2007) is the better choice (Farajian et al., 2017) for comparing quality of segmentation in our setting. DESEG consistently outperforms unsupervised alternatives BPE and MORFESSOR in Table 4 while approaching and even beating state-of-the-art systems FARASA and MADAMIRA in the individual–pooled environment. The Fertility-based model DESEG$_{f2}$ outperforms its greedy counterpart, supporting the argument that base fertility plays a meaningful role in morphological organization.

5 Error Analysis

We performed a quantitative error analysis on 100 sentences randomly selected from CORPUS6 for each variety, creating a *gold* segmentation set. In Table 5, accuracy is computed given the two modes of training DESEG$_{f2}$ (i.e., pooled or individual), and compared with the PLAIN input baseline. Average segmentation accuracy over all varieties correlates with the extrinsic evaluation for both modes of training DESEG$_{f2}$. In both modes, the best performance is on MSA and the worst is on Rabat then Tunis.

In individual mode, the poor performance of Rabat and Tunis is expected as we could not obtain sufficiently large monolingual data sets that distinguish these two quite linguistically distinct North African varieties. Thus, we were forced to train both grammars' disambiguators on the same data, propagating error whenever a form occurred in the Rabat dialect not analyzable by the Tunis grammar or vice versa. As for pooled mode, careful inspection revealed an exceptional amount of inconsistent spellings in the Tunis and Rabat partitions of CORPUS6 that were not anticipated when constructing the grammar. The definite article proclitic +ال *Al*+ for example, frequently appears as its own word, reduced to just ل *l*, or deleted altogether when preceded by another proclitic, especially when the ل *l* assimilates phonologically to the following phoneme. In MSA, by contrast, the definite article is always attached to the following noun, the ل *l* is never deleted, and the ا *A* can only be deleted following the prepositional proclitic +ل *l*+, 'for'. It is not surprising then that MSA performs the best in both modes as there is only negligible inconsistency in MSA spelling, meaning that the grammar need not anticipate an unbounded set of spelling alternatives exacerbating over generation and putting more stress on the disambiguator.

The best DA performance is achieved on Beirut for the pooled mode and Doha for the Individual. Beirut is the least verbose of all dialects in unsegmented space, and also exhibits the lowest ratio of unsegmented tokens to gold segmented tokens, meaning that it rewards over segmenting, which we know DESEG$_{f2}$ is biased toward given its secondary preference for short bases. As for the high performance on Doha, it is worth noting that Doha is also the highest performing dialect on all MT experiments, even recording higher BLEU scores than MSA. It is thus likely that the Doha partition of CORPUS6 is simply more internally consistent than the others, not just in terms of spelling, but also lexical choices and syntactic structure. This could be idiosyncratic to CORPUS6 more than it is characteristic of the Doha dialect, though an independent test corpus would be needed to investigate this further.

While the extrinsic MT results vouch for the effectiveness of pooled grammars when training data cannot be separated by dialect, the pooled training mode consistently fails to outperform PLAIN on the harsh evaluation metric of segmentation accuracy. On average, the pooled mode is 15% less accurate than individual—which does consistently improve over PLAIN—demonstrating that reducing the grammar's capacity to over generate by determining the dialect before segmenting greatly facilitates disambiguation. Indeed, there is a 94% correlation between the verbosity reduction and accuracy increase going from the pooled to individual mode, indicating that the pooled model is over segmenting as more options for mistakenly identifying segmentable clitics become available across different dialects.

This is especially problematic for words like the noun فرد *frd*, 'individual', which contain highly fertile, analyzable bases within their true base. That is, فرد *frd* can also be analyzed out of context as a conjunction followed by a verb ف+ رد *f+ rd* 'so he responded', where the verbal base رد is highly fertile, especially since it is identical to the nominal رد *rd*, 'response' and thus can participate in a large number of clitic combinations as licensed by three feasible meta paradigms (verbal PV, verbal CV, or nominal). Furthermore, the increased uncertainty caused by greater over generation of the analyzer in pooled mode gives the base length minimization back off more influence. Base length minimization as a disambiguation strategy will always over segment by definition if the analyzer permits it. Thus, low frequency or unknown words like the proper name اونو *Awnw*, 'Ono' are frequently over segmented, as occurs in all dialects except Doha and MSA, where the leading or trailing sequences of graphemes happen to not be confusable with any viable clitics according to the grammar.

Considering context will be crucial to improving the model's handling of such cases in future work, as the Cairene sentence هي دي مدام اونو؟ *hy dy mdAm Awnw?*, 'Is this Madame Ono?' provides a blatant clue in the title 'Madame', that اونو *Awnw* is a name and need not be segmented. Similarly, the Beiruti sentence, ازا بتريد صرفلي هالعشرة دولار عخمسة فرد شأفة ... *AzA btryd Srfly hAlςšrħ dwlAr ςxmsħ frd šÂfħ ...*, 'Please exchange for me this ten dollar [bill] for a single five...' indicates that a noun should follow the numerical modifier خمسة *xmsħ*, 'five', not the proclitic conjunction ف+ *f+*, 'so'.

	Seg Verbosity			Accuracy			Best
	Input	Pooled	Indiv	Input	Pooled	Indiv	ER
Beirut	0.69	1.22	1.13	56.7	68.7	**79.7**	53
Cairo	0.80	1.29	1.15	77.8	65.9	**81.3**	16
Rabat	0.72	1.30	1.19	66.1	57.9	**70.0**	11
Tunis	0.81	1.32	1.15	**79.4**	62.9	78.5	0
Doha	0.79	1.27	1.11	77.3	67.6	**85.2**	35
MSA	0.80	1.24	1.07	76.3	69.6	**88.3**	50
Average	0.77	1.27	1.13	72.3	65.4	**80.5**	30

Table 5: Segmentation accuracy of DESEG trained on Pooled versus Indiv(idual) dialects/grammars and evaluated on CORPUS6 against the PLAIN input baseline. Seg(mentation) verbosity is the ratio of segmented tokens over gold segmented tokens while accuracy and error reduction (ER) are reported as percentages.

6 Conclusion and Future Work

We present an effective unsupervised means of introducing linguistic information for segmentation that greatly improves performance over other unsupervised systems as evaluated both intrinsically and extrinsically. We target robust handling of rich morphological phenomena and noisy corpora, achieving performance on a multi-dialect Arabic corpus comparable to state-of-the-art supervised systems. The success of our simple system is strong evidence for the value of linguistic input during preprocessing.

In the future, we plan to evaluate our models on natural (uncommissioned) dialectal corpora. We also plan to enhance our delexicalize models with non-concatenative components. And we also in tend to develop models that consider context.

Acknowledgments

We would like to thank three anonymous reviewers for their feedback. This publication was made possible by grant NPRP 7-290-1-047 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A Fast and Furious

Segmenter for Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 11–16, San Diego, California.

Farrell Ackerman and Robert Malouf. 2013. Morphological Organization: The Low Conditional Entropy Conjecture. *Language*, 89(3):429–464.

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and San'ani Yemeni Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Shabib AlGahtani and John McNaught. 2015. Joint Arabic segmentation and part-of-speech tagging. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China.

Khalid Almeman and Mark Lee. 2013. Automatic building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words. In *Proceedings of the International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Faisal Alshargi, Shahd Dibas, Sakhar Alkhereyf, Reem Faraj, Basmah Abdulkareem, Sane Yagi, Ouafaa Kacha, Nizar Habash, and Owen Rambow. 2019. Morphologically Annotated Corpora for Seven Arabic Dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In *Proceedings of the Workshop on Arabic Natural Language Processing*, Florence, Italy.

Duygu Ataman and Marcello Federico. 2018. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. *arXiv preprint arXiv:1805.02036*.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331–342.

Mohammed Attia. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *Proceedings of the Conference on the Challenge of Arabic for NLP/MT*, London.

Mohammed Attia. 2007. Arabic Tokenization System. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL): Common Issues and Resources*, pages 65–72.

Kenneth Beesley. 1998. Arabic morphology using only finite-state operations. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*, pages 50–7, Montereal.

Raymond Bertram, R Harald Baayen, and Robert Schreuder. 2000. Effects of Family Size for Complex Words. *Journal of memory and language*, 42(3):390–405.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Mohamed Boudchiche, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja, and Abderrahim Boudlal. 2017. AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. *Journal of King Saud University - Computer and Information Sciences*, 29(2):141–146.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Mauro Cettolo and Christian Girardi. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Kareem Darwish. 2002. Building a Shallow Arabic Morphological Analyzer in One Day. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*, pages 47–54, Philadelphia, PA, USA.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 149–152, Boston, MA.

Markus Dreyer and Jason Eisner. 2011. Discovering Morphological Paradigms from Plain Text Using A Dirichlet Process Mixture Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 616–627, Edinburgh, United Kingdom.

Mohamed Eldesouki, Younes Samih, Ahmed Abdelali, Mohammed Attia, Hamdy Mubarak, Kareem Darwish, and Kallmeyer Laura. 2017. Arabic Multi-Dialect Segmentation: bi-LSTM-CRF vs. SVM. *arXiv preprint arXiv:1708.05891*.

Alexander Erdmann and Nizar Habash. 2018. Complementary Strategies for Low Resourced Morphological Modeling. In *Proceedings of the Workshop on Computational Research in Phonetics, Phonology, and Morphology (SIGMORPHON)*, pages 54–65, Brussels, Belgium.

Alexander Erdmann, Nizar Habash, Dima Taji, and Houda Bouamor. 2017. Low Resourced Machine Translation via Morpho-syntactic Modeling: The Case of Dialectal Arabic. In *Proceedings of the Machine Translation Summit (MT Summit)*.

Alexander Erdmann, Nasser Zalmout, and Nizar Habash. 2018. Addressing noise in multidialectal word embeddings. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Ramy Eskander, Nizar Habash, Owen Rambow, and Arfath Pasha. 2016a. Creating resources for Dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 3455–3465, Osaka, Japan.

Ramy Eskander, Owen Rambow, and Smaranda Muresan. 2018. Automatically Tailoring Unsupervised Morphological Segmentation to the Language. In *Proceedings of the Workshop of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, pages 78–83.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016b. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 900–910.

M Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. phrase-based machine translation in a multi-domain scenario. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page 280, Valencia, Spain.

Charles F Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Abed Alhakim Freihat, Gabor Bella, Hamdy Mubarak, and Fausto Giunchiglia. 2018. A Single-Model Approach for Arabic Segmentation, POS Tagging, and Named Entity Recognition. In *International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–8. IEEE.

Jin Guo. 1997. Critical tokenization and its properties. *Computational Linguistics*, 23(4):569–596.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Workshop of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, pages 1–9, Montréal, Canada.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 573–580, Ann Arbor, Michigan.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the International Conference on Computational Linguistics and the Conference of the Association for Computational Linguistics (COLING-ACL)*, pages 681–688, Sydney, Australia.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Atlanta, Georgia.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Hany Hassan, Mostafa Elaraby, and Ahmed Tawfik. 2017. Synthetic Data for Neural Machine Translation of Spoken-Dialects. *arXiv preprint arXiv:1707.00079*.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Katharina Kann, Stanislas Lauly, and Kyunghyun Cho. 2018. The NYU System for the CoNLL–SIGMORPHON 2018 Shared Task on Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 58–63, Brussels. Association for Computational Linguistics.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Salam Khalifa, Sara Hassan, and Nizar Habash. 2017. A Morphological Analyzer for Gulf Arabic Verbs. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Valencia, Spain.

Shereen Khoja and Roger Garside. 1999. Stemming Arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

Amit Kirschenbaum. 2015. To Split or Not, and If so, Where? Theoretical and Empirical Aspects of Unsupervised Morphological Segmentation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 139–150. Springer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language Model Based Arabic Word Segmentation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 399–406.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of the Workshop on*

Computational Approaches to Arabic Script-based Languages (CAASL), pages 2–9, Geneva, Switzerland.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *arXiv preprint arXiv:1503.02335*.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

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 (HLT–NAACL)*, pages 209–217, Boulder, Colorado.

Aleya Rouchdy. 2013. Language Conflict and Identity: Arabic in the American Diaspora. In *Language Contact and Language Conflict in Arabic*, pages 151–166. Routledge.

Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the International Conference on Computational Linguistics and the Conference of the Association for Computational Linguistics (COLING-ACL)*, pages 1–8, Sydney, Australia.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1332–1344, Santa Fe, New Mexico, USA.

Wael Salloum and Nizar Habash. 2014. ADAM: Analyzer for Dialectal Arabic Morphology. *Journal of King Saud University - Computer and Information Sciences*, 26(4):372–378.

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, and Laura Kallmeyer. 2017. Learning from Relatives: Unified Dialectal Arabic Segmentation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 432–441.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 1715–1725, Berlin, Germany.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-Supervised Morphological Segmentation using Adaptor Grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

Noah A. Smith and Jason Eisner. 2005. Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 354–362, Ann Arbor, Michigan.

Otakar Smrž. 2007. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague, Prague, Czech Republic.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 737–745, Columbus, Ohio.

David Stallard, Jacob Devlin, Michael Kayser, Yoong Keok Lee, and Regina Barzilay. 2012. Unsupervised Morphology Rivals Supervised Morphology for Arabic MT. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 322–327, Jeju Island, Korea.

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 *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 147–152, Las Palmas, Spain.

Clara Vania and Adam Lopez. 2017. From Characters to Words to in Between: Do We Capture Morphology? In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada. Association for Computational Linguistics.

Nasser Zalmout, Alexander Erdmann, and Nizar Habash. 2018. Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Orleans, Louisiana, USA.

Nasser Zalmout and Nizar Habash. 2017. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 704–713, Copenhagen, Denmark.

Nasser Zalmout and Nizar Habash. 2019. Adversarial Multitask Learning for Joint Multi-Feature and Multi-Dialect Morphological Modeling. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy.