Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French

Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi, Josef van Genabith

Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French

Mohammed Attia, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi, Josef van Genabith*
National Centre for Language Technology
School of Computing, Dublin City University
{mattia,jfoster,dhogan,jleroux,ltounsi,josef}@computing.dcu.ie

Abstract

This paper presents a study of the impact of using simple and complex morphological clues to improve the classification of rare and unknown words for parsing. We compare this approach to a language-independent technique often used in parsers which is based solely on word frequencies. This study is applied to three languages that exhibit different levels of morphological expressiveness: Arabic, French and English. We integrate information about Arabic affixes and morphotactics into a PCFG-LA parser and obtain state-of-the-art accuracy. We also show that these morphological clues can be learnt automatically from an annotated corpus.

1 Introduction

For a parser to do a reasonable job of analysing free text, it must have a strategy for assigning part-of-speech tags to words which are not in its lexicon. This problem, also known as the problem of unknown words, has received relatively little attention in the vast literature on Wall-Street-Journal (WSJ) statistical parsing. This is likely due to the fact that the proportion of unknown words in the standard English test set, Section 23 of the WSJ section of Penn Treebank, is quite small. The problem manifests itself when the text to be analysed comes from a different domain to the text upon which the parser has been trained, when the treebank upon which the parser has been trained is limited in size and when the language to be parsed is heavily inflected. We concentrate on the latter case, and examine the problem of unknown words for two languages which lie on opposite ends of the spectrum of morphological expressiveness and for one language which lies somewhere in between: Arabic, English and French.

In our experiments we use a Berkeley-style latent-variable PCFG parser and we contrast two techniques for handling unknown words within the generative parsing model: one in which no language-specific information is employed and one in which morphological clues (or signatures) are exploited. We find that the improvement accrued from looking at a word's morphology is greater for Arabic and French than for English. The morphological clues we use for English are taken directly from the Berkeley parser (Petrov et al., 2006) and those for French from recent work on French statistical parsing with the Berkeley parser (Crabbé and Candito, 2008; Candito et al., 2009). For Arabic, we present our own set of heuristics to extract these signatures and demonstrate a statistically significant improvement of 3.25% over the baseline model which does not employ morphological information.

We next try to establish to what extent these clues can be learnt automatically by extracting affixes from the words in the training data and ranking these using information gain. We show that this automatic method performs quite well for all three languages.

The paper is organised as follows: In Section 2 we describe latent variable PCFG parsing models. This is followed in Section 3 by a description of our three datasets, including statistics on the extent of the unknown word problem in each. In Section 4, we

*Author names are listed in alphabetical order. For further correspondence, contact L. Tounsi, D. Hogan or J. Foster.

Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 67–75,
Los Angeles, California, June 2010.

present results on applying a version of the parser which uses a simple, language-agnostic, unknown-word handling technique to our three languages. In Section 5, we show how this technique is extended to include morphological information and present parsing results for English and French. In Section 6, we describe the Arabic morphological system and explain how we used heuristic rules to cluster words into word-classes or signatures. We present parsing results for the version of the parser which uses this information. In Section 7, we describe our attempts to automatically determine the signatures for a language and present parsing results for the three languages. Finally, in Section 8, we discuss how this work might be fruitfully extended.

2 Latent Variable PCFG Parsing

Johnson (1998) showed that refining treebank categories with parent information leads to more accurate grammars. This was followed by a collection of linguistically motivated propositions for manual or semi-automatic modifications of categories in treebanks (Klein and Manning, 2003). In PCFG-LAs, first introduced by Matsuzaki *et al.* (2005), the refined categories are learnt from the treebank using unsupervised techniques. Each base category – and this includes part-of-speech tags – is augmented with an annotation that refines its distributional properties.

Following Petrov *et al.* (2006) latent annotations and probabilities for the associated rules are learnt incrementally following an iterative process consisting of the repetition of three steps.

1. Split each annotation of each symbol into n (usually 2) new annotations and create rules with the new annotated symbols. Estimate[1] the probabilities of the newly created rules.
2. Evaluate the impact of the newly created annotations and discard the least useful ones. Re-estimate probabilities with the new set of annotations.
3. Smooth the probabilities to prevent overfitting.

We use our own parser which trains a PCFG-LA using the above procedure and parses using the max-rule parsing algorithm (Petrov et al., 2006; Petrov and Klein, 2007). PCFG-LA parsing is relatively language-independent but has been shown to be very effective on several languages (Petrov, 2009). For our experiments, we set the number of iterations to be 5 and we test on sentences less than or equal to 40 words in length. All our experiments, apart from the final one, are carried out on the development sets of our three languages.

3 The Datasets

Arabic We use the the Penn Arabic Treebank (ATB) (Bies and Maamouri, 2003; Maamouri and Bies., 2004). The ATB describes written Modern Standard Arabic newswire and follows the style and guidelines of the English Penn-II treebank. We use the part-of-speech tagset defined by Bikel and Bies (Bikel, 2004). We employ the usual treebank split (80% training, 10% development and 10% test).

English We use the Wall Street Journal section of the Penn-II Treebank (Marcus et al., 1994). We train our parser on sections 2-21 and use section 22 concatenated with section 24 as our development set. Final testing is carried out on Section 23.

French We use the French Treebank (Abeillé et al., 2003) and divide it into 80% for training, 10% for development and 10% for final results. We follow the methodology defined by Crabbé and Candito (2008): compound words are merged and the tagset consists of base categories augmented with morphological information in some cases[2].

Table 1 gives basic unknown word statistics for our three datasets. We calculate the proportion of words in our development sets which are unknown or rare (specified by the cutoff value) in the corresponding training set. To control for training set size, we also provide statistics when the English training set is reduced to the size of the Arabic and French training sets and when the Arabic training set is reduced to the size of the French training set. In an ideal world where training set sizes are the same for all languages, the problem of unknown words will be greatest for Arabic and smallest for English. It is

[1] Estimation of the parameters is performed by running Expectation/Maximisation on the training corpus.

[2] This is called the CC tagset: base categories with verbal moods and extraction features

language	cutoff	#train	#dev	#unk	%unk	language	#train	#dev	#unk	%unk
Arabic	0	594,683	70,188	3794	5.40	Reduced English	597,999	72,970	2627	3.60
-	1	-	-	6023	8.58	(Arabic Size)	-	-	3849	5.27
-	5	-	-	11,347	16.17	-	-	-	6700	9.18
-	10	-	-	15,035	21.42	-	-	-	9083	12.45
English	0	950,028	72,970	2062	2.83	Reduced Arabic	266,132	70,188	7027	10.01
-	1	-	-	2983	4.09	(French Size)	-	-	10,208	14.54
-	5	-	-	5306	7.27	-	-	-	16,977	24.19
-	10	-	-	7230	9.91	-	-	-	21,434	30.54
French	0	268,842	35,374	2116	5.98	Reduced English	265,464	72,970	4188	5.74
-	1	-	-	3136	8.89	(French Size)	-	-	5894	8.08
-	5	-	-	5697	16.11	-	-	-	10,105	13.85
-	10	-	-	7584	21.44	-	-	-	13,053	17.89

Table 1: Basic Unknown Word Statistics for Arabic, French and English

reasonable to assume that the levels of inflectional richness have a role to play in these differences.

4 A Simple Lexical Probability Model

The simplest method for handling unknown words within a generative probabilistic parsing/tagging model is to reserve a proportion of the lexical rule probability mass for such cases. This is done by mapping rare words in the training data to a special UNKNOWN terminal symbol and estimating rule probabilities in the usual way. We illustrate the process with the toy unannotated PCFG in Figures 1 and 2. The lexical rules in Fig. 1 are the original rules and the ones in Fig. 2 are the result of applying the rare-word-to-unknown-symbol transformation. Given the input sentence *The shares recovered*, the word *recovered* is mapped to the UNKNOWN token and the three edges corresponding to the rules $NNS \rightarrow$ UNKNOWN, $VBD \rightarrow$ UNKNOWN and $JJ \rightarrow$ UNKNOWN are added to the chart at this position. The disadvantage of this simple approach is obvious: all unknown words are treated equally and the tag whose probability distribution is most dominated by rare words in the training will be deemed the most likely (JJ for this example), regardless of the characteristics of the individual word. Apart from its ease of implementation, its main advantage is its language-independence - it can be used off-the-shelf for any language for which a PCFG is available.[3]

One parameter along which the simple lexical probability model can vary is the threshold used to decide whether a word in the training data is rare or "unknown". When the threshold is set to *n*, a word in the training data is considered to be unknown if it occurs *n* or fewer times. We experiment with three thresholds: 1, 5 and 10. The result of this experiment for our three languages is shown in Table 2.

The general trend we see in Table 2 is that the number of training set words considered to be unknown should be minimized. For all three languages, the worst performing grammar is the one obtained when the threshold is increased to 10. This result is not unexpected. With this simple lexical probability model, there is a trade-off between obtaining good guesses for words which do not occur in the training data and obtaining reliable statistics for words which do. The greater the proportion of the probability mass that we reserve for the unknown word section of the grammar, the more performance suffers on the known yet rare words since these are the words which are mapped to the UNKNOWN symbol. For example, assume the word *restructuring* occurs 10 times in the training data, always tagged as a *VBG*. If the unknown threshold is less than ten and if the word occurs in the sentence to be parsed, a *VBG* edge will be added to the chart at this word's position with the probability 10/#VBG. If, however, the threshold is set to 10, the word (in the training set and the input sentence) will be mapped to UNKNOWN and more possibilities will be explored (an edge for each $TAG \rightarrow$ UNKNOWN rule in the grammar). We can see from Table 1 that at threshold 10, one fifth

[3] Our simple lexical model is equivalent to the Berkeley simpleLexicon option.

```
VBD -> fell 50/153
VBD -> reoriented 2/153
VBD -> went 100/153
VBD -> latched 1/153
NNS -> photofinishers 1/201
NNS -> shares 200/201
JJ  -> financial 20/24
JJ  -> centrist 4/24
DT  -> the 170/170
```

Figure 1: The original toy PCFG

```
VBD -> fell 50/153
VBD -> UNKNOWN 3/153
VBD -> went 100/153
NNS -> UNKNOWN 1/201
NNS -> shares 200/201
JJ  -> financial 20/24
JJ  -> UNKNOWN 4/24
DT  -> the 170/170
```

Figure 2: Rare → UNKNOWN

```
VBD -> fell 50/153
VBD -> UNK-ed 3/153
VBD -> went 100/153
NNS -> UNK-s 1/201
NNS -> shares 200/201
JJ  -> financial 20/24
JJ  -> UNK-ist 4/24
DT  -> the 170/170
```

Figure 3: Rare → UNKNOWN+SIGNATURE

Unknown Threshold	Recall	Precision	F-Score	Tagging Accuracy
	Arabic			
1	78.60	80.49	**79.53**	94.03
5	77.17	79.81	78.47	91.16
10	75.32	78.69	76.97	89.06
	English			
1	89.20	89.73	**89.47**	95.60
5	88.91	89.74	89.33	94.66
10	88.00	88.97	88.48	93.61
	French			
1	83.60	84.17	**83.88**	94.90
5	82.31	83.10	82.70	92.99
10	80.87	82.05	81.45	91.56

Table 2: Varying the Unknown Threshold with the Simple Lexical Probability Model

of the words in the Arabic and French development sets are unknown, and this is reflected in the drop in parsing performance at these thresholds.

5 Making use of Morphology

Unknown words are not all the same. We exploit this fact by examining the effect on parsing accuracy of clustering rare training set words using cues from the word's morphological structure. Affixes have been shown to be useful in part-of-speech tagging (Schmid, 1994; Tseng et al., 2005) and have been used in the Charniak (Charniak, 2000), Stanford (Klein and Manning, 2003) and Berkeley (Petrov et al., 2006) parsers. In this section, we contrast the effect on parsing accuracy of making use of such information for our three languages of interest.

Returning to our toy English example in Figures 1 and 2, and given the input sentence *The shares recovered*, we would like to use the fact that the unknown word *recovered* ends with the past tense suffix *-ed* to boost the probability of the lexical rule $VBD \rightarrow$ *UNKNOWN*. If we specialise the `UNKNOWN` terminal using information from English morphology, we can do just that, resulting in the grammar in Figure 3. Now the word *recovered* is mapped to the symbol `UNK-ed` and the only edge which is added to the chart at this position is the one corresponding to the rule $VBD \rightarrow$ `UNK-ed`.

For our English experiments we use the unknown word classes (or *signatures*) which are used in the Berkeley parser. A signature indicates whether a words contains a digit or a hyphen, if a word starts with a capital letter or ends with one of the following English suffixes (both derivational and inflectional): *-s*, *-ed*, *-ing*, *-ion*, *-er*, *-est*, *-ly*, *-ity*, *-y* and *-al*.

For our French experiments we employ the same signature list as Crabbé and Candito (2008), which itself was adapted from Arun and Keller (2005). This list consists of (a) conjugation suffixes of regu-

lar verbs for common tenses (eg. *-ons, -ez, -ent*...) and (b) derivational suffixes for nouns, adverbs and adjectives (eg. *-tion, -ment, -able*...).

The result of employing signature information for French and English is shown in Table 3. Beside each f-score the absolute improvement over the UNKNOWN baseline (Table 2) is given. For both languages there is an improvement at all unknown thresholds. The improvement for English is statistically significant at unknown thresholds 1 and 10.[4] The improvement is more marked for French and is statistically significant at all levels.

In the next section, we experiment with signature lists for Arabic.[5]

6 Arabic Signatures

In order to use morphological clues for Arabic we go further than just looking at suffixes. We exploit all the richness of the morphology of this language which can be expressed through morphotactics.

6.1 Handling Arabic Morphotactics

Morphotactics refers to the way morphemes combine together to form words (Beesley, 1998; Beesley and Karttunen, 2003). Generally speaking, morphotactics can be concatenative, with morphemes either prefixed or suffixed to stems, or non-concatenative, with stems undergoing internal alternations to convey morphosyntactic information. Arabic is considered a typical example of a language that employs non-concatenative morphotactics.

Arabic words are traditionally classified into three types: verbs, nouns and particles. Adjectives take almost all the morphological forms of, and share the same templatic structures with, nouns. Adjectives, for example, can be definite, and are inflected for case, number and gender.

There are a number of indicators that tell us whether the word is a verb or a noun. Among these indicators are prefixes, suffixes and word templates. A template (Beesley and Karttunen, 2003) is a kind of vocalization mould in which a word fits. In derivational morphology Arabic words are formed through the amalgamation of two tiers, namely, root and template. A root is a sequence of three (rarely two or four) consonants which are called radicals, and the template is a pattern of vowels, or a combination of consonants and vowels, with slots into which the radicals of the root are inserted.

For the purpose of detection we use the reverse of this information. Given that we have a word, we try to extract the stem, by removing prefixes and suffixes, and match the word against a number of verbal and nominal templates. We found that most Arabic templatic structures are in complementary distribution, i.e. they are either restricted to nominal or verbal usage, and with simple regular expression matching we can decide whether a word form is a noun or a verb.

6.2 Noun Indicators

In order to detect that a word form is a noun (or adjective), we employ heuristic rules related to Arabic prefixes/suffixes and if none of these rules apply we attempt to match the word against templatic structures. Using this methodology, we are able to detect 95% of ATB nouns.[6]

We define a list of 42 noun templates which are used to indicate active/passive participle nouns, verbal nouns, nouns of instrument and broken plural nouns (see Table 4 for some examples). Note that templates ending with taa marboutah "ap" or starting with meem madmoumah "mu" are not considered since they are covered by our suffix/prefix rules, which are as follows:

1- The definite article prefix ال or in Buckwalter transliteration "Al".

2- The tanween suffix اً, اٍ, اٌ or "N", "F", "K", "AF".

3- The feminine plural suffix ات, or "+At".

4- The taa marboutah ending ة or "ap" whether as a

[4] Statistical significance was determined using the stratified shuffling method. The software used to perform the test was downloaded from http://www.cis.upenn.edu/~dbikel/software.html.

[5] An inspection of the Berkeley Arabic grammar (available at http://code.google.com/p/berkeleyparser/downloads/list) shows that no Arabic-specific signatures were employed. The Stanford parser uses 9 signatures for Arabic, designed for use with unvocalised text. An immediate future goal is to test this signature list with our parser.

[6] The heuristics we developed are designed to work on diacritized texts. Although diacritics are generally ignored in modern writing, the issue of restoring diacritics has been satisfactorily addressed by different researchers. For example, Nelken and Shieber (2005) presented an algorithm for restoring diacritics to undiacritized MSA texts with an accuracy of over 90% and Habasah *et al.* (2009) reported on a freely-available toolkit (MADA-TOKAN) an accuracy of over 96%.

Unknown Threshold	Recall	Precision	F-Score	Tagging Accuracy
	Arabic			
1	80.67	82.19	*81.42 (+ 1.89)	96.32
5	80.66	82.81	*81.72 (+ 3.25)	95.15
10	79.86	82.49	*81.15 (+ 4.18)	94.38
	English			
1	***89.64	89.95	**89.79** (+ 0.32)	96.44
5	89.16	89.80	89.48 (+ 0.15)	96.32
10	89.14	89.78	**89.46 (+ 0.98)	96.21
	French			
1	85.15	85.77	***85.46 (+ 1.58)**	96.13
5	84.08	84.80	*84.44 (+ 1.74)	95.54
10	84.21	84.78	*84.49 (+ 3.04)	94.68

Table 3: Baseline Signatures for Arabic, French and English
statistically significant with *:$p < 10^{-4}$, **: $p < 10^{-3}$, ***: $p < 0.004$,

Template Name		Regular	Specification
Arabic	Buckwalter	Expression	
إِنفِعَال	{inofiEAl	{ino.i.A.	verbal noun (masdar)
مِفعَال	mifoEAl	mi.o.A.	noun instrument
مُستَفعِل	musotafoEil	musota.o.i.	noun participle
مَفَاعِيل	mafAEiyl	ma.A.iy.	noun plural
إِستَفعَل	{isotafoEal	{isota.o.a.	verb
فُوعِل	fuwEil	.uw.i.	verb passive

Table 4: Sample Arabic Templatic Structures for Nouns and Verbs

feminine marker suffix or part of the word.
5- The genitive case marking kasrah إ or "+i".
6- Words of length of at least five characters ending with doubled yaa يّ or "y~".
7- Words of length of at least six characters ending with alif mamdoudah and hamzah اء or "A'".
8- Words of length of at least seven characters starting with meem madmoumah مُ or "mu".

6.3 Verb Indicators

In the same way, we define a list of 16 templates and we combine them with heuristic rules related to Arabic prefixes/suffixes to detect whether a word form is exclusively a verb. The prefix/suffix heuristics are as follows:
9-The plural marker suffix وَا or "uwA" indicates a verb.
10- The prefixes ت ،ي ،ن ،أَ ،سَ or "sa", ">a", ">u", "na", "nu", "ya", "yu", "ta", "tu" indicate imprefective verb.
The verbal templates are less in number than the noun templates yet they are no less effective in detecting the word class (see Table 4 for examples). Using these heuristics we are able to detect 85% of ATB verbs.

6.4 Arabic Signatures

We map the 72 noun/verb classes that are identified using our hand-crafted heuristics into sets of signatures of varying sizes: 4, 6, 14, 21, 25, 28 and 72. The very coarse-grained set considers just 4 signatures `UNK-noun`, `UNK-verb`, `UNK-num`, and `UNK` and the most fine-grained set of 72 signatures associates one signature per heuristic. In addition, we have evaluated the effect of reordering rules and templates and also the effect of collating all signatures satisfying an unknown word. The results of using these various signatures sets in parsing

UNK					
NUM digits	**NOUN** (see section 6.2)				**VERB** (see section 6.3)
	Al_definiteness rule 1	**tashkil** rules 2 and 5	**At_suffix** rule 3	**ap_suffix** rule 4	**imperfect** rule 10
	y¯_suffix rule 6	**A′_suffix** rule 7	**mu_prefix** rule 8	**verbal_noun_templates** 3 groupings	**suffixes** dual/plural suffixes
	plural_templates 4 groupings	**participle_active_templates**	**participle_passive_templates**	**instrument_templates**	**passive_templates**
	other_templates				**verbal templates** 5 groupings

Table 6: Arabic signatures

Cutoff	**1**	**5**	**10**
4	80.78	80.71	80.09
6	81.14	81.16	81.06
14	80.88	81.45	81.19
14 reorder	81.39	81.01	80.81
21	81.38	81.55	81.35
21 reorder	81.20	81.13	80.58
21 collect	80.94	80.56	79.63
25	81.18	81.25	81.26
28	81.42	**81.72 (+ 3.25)**	81.15
72	79.64	78.87	77.58

Table 5: Baseline Signatures for Arabic

our Arabic development set are presented in Table 5. We achieve our best labeled bracketing f-score using 28 signatures with an unknown threshold of five. In fact we get an improvement of 3.25% over using no signatures at all (see Table 2). Table 3 describes in more detail the scores obtained using the 28 signatures present in Table 6. Apart from the set containing 72 signatures, all of the baseline signature sets in Table 5 yield a statistically significant improvement over the generic UNKNOWN results ($p < 10^{-4}$).

7 Using Information Gain to Determine Signatures

It is clear that dividing the UNKNOWN terminal into more fine-grained categories based on morphological information helps parsing for our three languages. In this section we explore whether useful morphological clues can be learnt automatically. If they can, it means that a latent-variable PCFG parser can be adapted to any language without knowledge of the language in question since the only language-specific component in such a parser is the unknown-signature specification.

In a nutshell, we extract affix features from training set words[7] and then use information gain to rank these features in terms of their predictive power in a POS-tagging task. The features deemed most discriminative are then used as signatures, replacing our baseline signatures described in Sections 5 and 6. We are not going as far as actual POS-tagging, but rather seeing whether the affixes that make good features for a part-of-speech tagger also make good unknown word signatures.

We experiment with English and French suffixes of length 1-3 and Arabic prefixes and suffixes of various lengths as well as stem prefixes and suffixes of length 2, 4 and 6. For each of our languages we experiment with several information gain thresholds on our development sets and we fix on an English signature list containing 24 suffixes, a French list containing 48 suffixes and an Arabic list containing 38 prefixes and suffixes.

Our development set results are presented in Table 7. For all three languages, the information gain signatures perform at a comparable level to the baseline hand-crafted signatures (Table 3). For each of the three unknown-word handling techniques, no signature (UNKNOWN), hand-crafted signatures and information gain signatures, we select the best unknown threshold for each language's development set and apply these grammars to our test sets. The f-scores are presented in Table 8, along with the upper bounds obtained by parsing with these grammars in gold-tag mode. For French, the effect of tagging accuracy on overall parse accuracy is striking. The improvements that we get from using morphological signatures are greatest for Arabic[8] and smallest for

[7] We omit all function words and high frequency words because we are interested in the behaviour of words which are likely to be similar to rare words.

[8] Bikel's parser trained on the same Arabic data and tested on the same input achieves an f-score of 76.50%. We trained a 5-split-merge-iteration Berkeley grammar and parsed with the

Unknown Threshold	Recall	Precision	F-Score	Tagging Accuracy
	Arabic IG			
1	80.10	82.15	*81.11 (+ 1.58)	96.53
5	80.03	82.49	***81.32 (+ 2.85)**	95.30
10	80.17	82.40	*81.27 (+ 4.3)	94.66
	English IG			
1	89.38	89.87	89.63 (+ 0.16)	96.45
5	89.54	90.22	***89.88 (+ 0.55)**	96.41
10	89.22	90.05	*89.63 (+ 1.15)	96.19
	French IG			
1	84.78	85.36	***85.07 (+ 1.19)**	96.17
5	84.63	85.24	**84.93 (+ 2.23)	95.30
10	84.18	84.80	*84.49 (+ 3.09)	94.68

Table 7: Information Gain Signature Results
statistically significant with *:$p < 10^{-4}$, **: $p < 2 \cdot 10^{-4}$, ***: $p < 0.005$

Language	No Sig	Baseline Sig	IG Sig
Arabic	78.34	*81.59	*81.33
Arabic Gold Tag	81.46	82.43	81.90
English	89.48	89.65	89.77
English Gold Tag	89.94	90.10	90.23
French	83.74	*85.77	**85.55
French Gold Tag	88.82	88.41	88.86

statistically significant with *: $p < 10^{-4}$, **: $p < 10^{-3}$

Table 8: F-Scores on Test Sets

English. The results for the information gain signatures are promising and warrant further exploration.

8 Conclusion

We experiment with two unknown-word-handling techniques in a statistical generative parsing model, applying them to Arabic, French and English. One technique is language-agnostic and the other makes use of some morphological information (signatures) in assigning part-of-speech tags to unknown words. The performance differences from the two techniques are smallest for English, the language with the sparsest morphology of the three and the smallest proportion of unknown words in its development set. As a result of carrying out these experiments, we have developed a list of Arabic signatures which can be used with any statistical parser which does its own tagging. We also present results which show that signatures can be learnt automatically.

Our experiments have been carried out using gold tokens. Tokenisation is an issue particularly for Arabic, but also for French (since the treebank contains merged compounds) and to a much lesser extent for English (unedited text with missing apostrophes). It is important that the experiments in this paper are repeated on untokenised text using automatic tokenisation methods (e.g. MADA-TOKAN).

The performance improvements that we demonstrate for Arabic unknown-word handling are obviously just the tip of the iceberg in terms of what can be done to improve performance on a morphologically rich language. The simple generative lexical probability model we use can be improved by adopting a more sophisticated approach in which known and unknown word counts are combined when estimating lexical rule probabilities for rare words (see Huang and Harper (2009) and the Berkeley sophisticatedLexicon training option). Further work will also include making use of a lexical resource external to the treebank (Goldberg et al., 2009; Habash, 2008) and investigating clustering techniques to reduce data sparseness (Candito and Crabbé, 2009).

Acknowledgements

This research is funded by Enterprise Ireland (CFTD/07/229 and PC/09/037) and the Irish Research Council for Science Engineering and Technology (IRCSET). We thank Marie Candito and our three reviewers for their very helpful suggestions.

Berkeley parser, achieving an f-score of 75.28%. We trained the Berkeley parser with the *-treebank SINGLEFILE* option so that English signatures were not employed.

References

Anne Abeillé, Lionel Clément, and François Toussenel, 2003. *Treebanks: Building and Using Parsed Corpora*, chapter Building a Treebank for French. Kluwer, Dordrecht.

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *ACL*. The Association for Computer Linguistics.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI studies in computational linguistics.

Kenneth R. Beesley. 1998. Arabic morphology using only finite-state operations. In *The Workshop on Computational Approaches to Semitic Languages*.

Ann Bies and Mohammed Maamouri. 2003. Penn Arabic Treebank guidelines. Technical Report TB-1-28-03.

Dan Bikel. 2004. *On the Parameter Space of Generative Lexicalized Parsing Models*. Ph.D. thesis, University of Pennslyvania.

Marie Candito and Benoit Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of IWPT'09*.

Marie Candito, Benoît Crabbé, and Djamé Seddah. 2009. On statistical parsing of French with supervised and semi-supervised strategies. In *Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference*, pages 49–57, Athens, Greece, March.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the Annual Meeting of the North American Association for Computational Linguistics (NAACL-00)*, pages 132–139, Seattle, Washington.

Benoît Crabbé and Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. In *Actes de TALN*.

Yoav Goldberg, Reut Tsarfaty, Meni Adler, and Michael Elhadad. 2009. Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and EM-HMM-based lexical probabilities. In *EACL*, pages 327–335. The Association for Computer Linguistics.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of Association for Computational Linguistics*, pages 57–60.

Zhongqiang Huang and Mary Harper. 2009. Self-training pcfg grammars with latent annotations across languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August.

Mark Johnson. 1998. PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

Dan Klein and Chris Manning. 2003. Accurate unlexicalised parsing. In *Proceedings of the 41st Annual Meeting of the ACL*.

Mohammed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, guidelines, procedures, and tools. In *Workshop on Computational Approaches to Arabic Script-based Languages, COLING*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the 1994 ARPA Speech and Natural Language Workshop*, pages 114–119, Princeton, New Jersey.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic CFG with latent annotations. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 75–82, Ann Arbor, June.

Rani Nelken and Stuart M. Shieber. 2005. Arabic diacritization using weighted finite-state transducers. In *ACL-05 Workshop on Computational Approaches to Semitic Languages*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*, Rochester, NY, April.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, Sydney, Australia, July.

Slav Petrov. 2009. *Coarse-to-Fine Natural Language Processing*. Ph.D. thesis, University of California at Berkeley, Berkeley, CA, USA.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP-1)*, pages 44–49.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.