INTERSPEECH 2009 BRIGHTON

Morphological Analysis and Decomposition for Arabic Speech-to-Text Systems

F. Diehl, M.J.F. Gales, M. Tomalin, & P.C. Woodland

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.

{fd257, mjfg, mt126, pcw}@eng.cam.ac.uk

Abstract

Language modelling for a morphologically complex language such as Arabic is a challenging task. Its agglutinative structure results in data sparsity problems and high out-of-vocabulary rates. In this work these problems are tackled by applying the MADA tools to the Arabic text. In addition to morphological decomposition, MADA performs context-dependent stem-normalisation. Thus, if word-level system combination, or scoring, is required this normalisation must be reversed. To address this, a novel context-sensitive method for morpheme-to-word conversion is introduced. The performance of the MADA decomposed system was evaluated on an Arabic broadcast transcription task. The MADA-based system out-performed the word-based system, with both the morphological decomposition and stem normalisation being found to be important.

Index Terms: Arabic, STT, MADA, morphology

10.21437/Interspeech.2009-122

1. Introduction

It is well-known that Arabic is a agglutinative, morphologically complex language which makes extensive use of clitics in order to convey information about parts-of-speech, gender, number, case, and so on. This causes non-trivial problems when Arabic speech is modelled in state-of-the-art Speech-to-Text (STT) systems. If morphological decomposition is not applied to the Arabic text, then any given word has many different morphologically distinct forms, and these must all be handled separately. For CallHome data, for instance, the vocabulary growth rate for Arabic is approximately 2.5 times higher than the rate for English [1]. This causes comparatively high Out-of-Vocabulary (OOV) rates if conventional modelling approaches are simply adapted from STT systems that were designed for non-agglutinative languages. For instance, for a 64K dictionary Afify *et al* report a 0.5% OOV rate for English but a 5% OOV rate for Arabic [2].

In recent years, various morphological decomposition schemes have been explored for Arabic STT systems trying to balance the advantage of a reduced OOV rate against the downside of a reduction in relative language model (LM) span by splitting words in morpheme sequences. Vergyri *et al* use a factored Language Model (LM) the factors of which (a root, a pattern, and a morphological class) are obtained by a morphological analyser [3]. Applying this LM in a medium sized multi-pass decoding task results in a WER reduction of 3.3% relative. Afify *et al* explore rule-based approaches to morphologically analysed dialectal Arabic speech recognition [2]. Pre-defined lists of 18 prefixes and 33 suffixes are applied to Arabic word tokens and simple rules separate the affixes. For a STT task that uses a 60K word list, WER reductions of 13% relative are reported. Xiang *et al* adopt a similar approach: Arabic words are divided into a prefix, a stem, and a suffix using pre-specified lists [4]. A complex decomposition algorithm is used: artificial compound words are introduced, and this gives WER reductions of 10% relative for a 64K word STT task. By contrast, Choueiter *et al* apply a morpheme generator to perform prefix and suffix decomposition. For a 64K word STT task, an 8% relative reduction in WER is obtained. For a STT task of 800K words this improvement reduces to 0.7% relative [5]. More recently, Lamel *et al* describe experiments for a 290K task. Though no improvements were obtained for a stand-alone system, reductions in WER of about 2%-5% relative were obtained using system combination [6].

A morphologically motivated prefix-stem decomposition which also includes a stem normalisation is provided by the 'D2' configuration of the 'Morphological Analysis and Disambiguation for Arabic' (MADA) tools[1]. These tools have successfully been used as a text pre-processing stage for Arabic-to-English Statistical Machine Translation (SMT) [7], [8], [9]. Though [10] reports on the use of MADA for Arabic dictionary generation, until now no work has been published on using MADA for OOV rate reduction by morphological decomposition. One reason for this is the stem normalisation property of the MADA tools which complicates the morpheme-to-word back-mapping.

This work examines how MADA-based decomposition can be applied to a state-of-the-art LVCSR Arabic STT system. As these systems are scored at the word-level, a morpheme-to-word back-mapping approach to handle the stem normalisation is required. A novel method for morpheme-to-word conversion motivated by N-gram SMT approaches is described. In addition, this back-mapping allows the combination of MADA-decomposed systems with word-based systems.

This paper is organised as follows: Section 2 introduces the MADA tools, and Section 3 discusses the morpheme-to-word conversion. Experimental results are given in 4.1 which is followed by the conclusions in Section 5.

2. Morphological Decomposition

Arabic is a highly inflected, morphologically rich language in which grammatical features, such as gender, number, person, and voice, are indicated by the attachment of clitics to lexical roots. The attachment of conjunction proclitics, particle proclitics, the definite article *Al*, and pronominal enclitics is largely rule-governed, and the basic patterns are indicated in the following schema [11]:

$$[CONJ+ \quad [PART+ \quad [Al+ \quad ROOT \quad +PRON]]]$$

The MADA tools can be used to obtain a morphological decomposition of the Arabic words. These tools implement a tokenisa-

This work was in part supported by DARPA under the GALE programme via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

[1]The MADA tools can be obtained as MADA_Distribution.html from http://www1.cs.columbia.edu/ƒambow/software-downloads.

tion and tagging stage which is then followed by a morphological disambiguation stage. The final output contains decomposed morpheme sequences [12].

The MADA tools support various levels of morphological detail. In the D2 configuration considered here, four particle proclitics ($l+$ (to/for), $b+$ $(by/with)$, $k+$ $(as/such)$, and $f+$ (in)) and one conjunction proclitic ($w+$ (and)) are identified. These are separated from their associated word roots [7].[2] For example, the complex structure '$lrjAl$' (literally, 'to a man') is decomposed into the morpheme sequence '$l+$ $rjAl$'. It is also possible to modify the basic D2 processing. For instance, the processing referred to as D2+Al in this paper separates the same morphemes as D2, but, in addition, it separates all definite articles (Al) which do not precede one of the so-called 'solar' consonants.[3] The D2+Al scheme has a linguistic motivation: the solar consonants alter the way in which the definite article Al is pronounced.

The following three schemes are discussed in this paper:

- **Word**: a word-based system, with no morphological decomposition (baseline);
- **D2**: the prefixes $l+$, $b+$, $k+$, $f+$, and $w+$ are separated from the word stem using MADA tools;
- **D2+Al**: the **D2** data is post-processed so that all Als which do not precede v t S \$ s z r d D Z C l n and T are separated from the word stem.

Examples for both decomposition schemes are given in Table 1. In addition to illustrating the different types of decomposition, these examples also indicate the stem normalisation performed by MADA: the second l in $wllREb$ becomes *Al*; the second *A* (*alif*) in *AlAyrAnY* becomes *M* (*alif with hamza below*), and the *Y* (*alif maksura*) in *AlAyrAnY* becomes *y* (*yeh*). As these examples indicate, due to the stem normalisation, MADA does not provide a bidirectional mapping from the word domain to the 'morpheme' domain. Typically, approximately 20% of the tokens are affected by stem normalisation.

System	Example sentence				
Word			*wll\$Eb*		*AlAyrAnY*
D2	*w+*	*l+*	*Al\$Eb*		*AlMyrAny*
D2+Al	*w+*	*l+*	*Al\$Eb*	*Al+*	*MyrAny*

Table 1: Examples of D2 and D2+Al morphologically decomposed data.

In this work prior MADA processing, a normalisation step to regularise a range of common inconsistencies of the Arabic orthography was applied. This involved mapping non-word-final *alif maksura* and all other *alif*s to plain *alif*; mapping word-final *yeh* to *alif maksura*; deleting the short vowel markers *fatha*, *damma*, *kasra*, *fathatan*, *dammatan*, *kasratan*; and deleting the vowel omission marker *sukun* and the consonant gemination marker *shadda*.

3. Morpheme-to-word Mapping

The MADA stem normalisation complicates the morpheme-to-word conversion. In contrast to other morphological decomposition schemes the prefixes cannot simply be reattached to their stems. Instead an additional stem back-mapping is necessary. The stem normalisation performed by MADA is however a context-sensitive process. Just applying a 1-gram based back-mapping according to a look-up table is not appropriate, rather a context dependent mapping procedure is needed.

To solve this problem, a novel approach was adopted in which the MADA-to-word conversion was viewed as a Statistical Machine Translation (SMT) task, and where the MADA-domain is the source 'language', while the word-domain is the target 'language'. Thus, the task involves 'translating' morpheme sequences into word sequences. This particular 'translation' problem involves many-to-one mappings from the source to the target, and many-to-one mappings from the target to the source do not occur. The source tokens were never mapped to null. Further, the approach does not require any reordering, and it gives a linear alignment between the source domain tokens and the target domain tokens.

Translation problems structured like this are well defined, and the N-gram SMT approach provides an efficient solutions to them [13]. N-gram SMT applies a 'bilingual' LM trained on so-called 'tuples'. These tuples (t, s), forming the basic bilingual units, assign to one or more source tokens s exactly one target token t. They are obtained by an alignment of the source and target data.

The translation model probability $p(T, S)$ from a source sentence S to a target sentence T is given by N-grams of tuples. For a sentence pair with K target tokens $p(T, S)$ is approximated by:

$$p(T,S) = \prod_{k=1}^{K} p((t,s)_k | (t,s)_{k-1}, ..., (t,s)_{k-N+1}) \qquad (1)$$

Decoding is done by likelihood maximisation of $p(T|S)$ with respect to $T = (t_1, t_2, ..., t_K)$ [14].

As a side-effect of the stem back-mapping, the N-gram SMT approach implicitly solves the problem of rejoining the split-off prefixes to their subsequent stems. This is effectively caused by the many-to-one assignments from MADA-domain tokens to word-domain tokens to form the tuples. With respect to the example given in Table 1, one of these tuples would be: (*wllREb*, *w+ l+ AlREb*), effectively rejoining the two split-off prefixes to their stem.

Due to unseen tuples in the training data of the bilingual language model, some MADA domain tokens can not be translated to the word domain. In such Out-of-Tuples (OOT) cases possible prefixes are joined to the stem and the resulting MADA domain word is directly transferred to the word domain. In cases where a normalised stem or a prefix is involved this may produce a new word which is not in the vocabulary. However, as only about 20% of the tokens (see section 2) are affected by stem normalisation, simply recomposing the word by joining the prefixes to the stem will yield a reasonable back-mapping. procedure.

4. Experiments and Results

4.1. System Description

STT systems were built using the word-based tokenisation of the training data, as well as D2 and D2+Al tokenisation. All acoustic models were trained on 1538 hours of acoustic data, and individual models for all word-based and morpheme-based systems were built. For development purposes Maximum Likelihood (ML) trained models were used, and for the final system evaluation, Minimum Phone Error (MPE) trained acoustic models were built. To facilitate the development of morpheme-based

[2] The Buckwalter romanised Arabic transliteration conventions are used throughout this paper.

[3] In Buckwalter notation the solar consonants are v t S \$ s z r d D Z C l n T.

systems, throughout this work the acoustic models were based on grapheme units. However for the final system evaluation, these graphemic system are combined with a phonetic word-based system.

All systems used an acoustic front-end which provided a 39-dimensional feature vector. Context-dependent cross-word and cross-morpheme Hidden Markov Models (HMM) were trained. The number of graphemic units was 36 whereas for the vowelised system 39 units (36 + 3 for short vowel modelling) were used. Trigrapheme models and triphone models, respectively, were produced via decision-tree state clustering, giving about 9000 distinct states with 36 Gaussian components per state on average. The development LMs were built using 8 sources which contained broadcast news, broadcast conversation, newswire, and web data. In total there were 522M words of LM training data. For the final evaluation systems approximately twice as much text data from 24 sources wee used. For all LMs the vocabulary size was 350K words or morphemes, respectively.

All the experiments in this section were based on a 3-pass system combination framework. For initial development using the ML-trained models, only the first two passes of the system, P1-P2, were used. This uses an initial rapid decoding first pass. The generated hypothesis is used to estimate least-squares-linear-regression and diagonal variance transforms. The adapted models are then used in a second, lattice-generation, pass with a trigram language model. These trigram lattices are expanded with a 4-gram LM and Confusion network (CN) decoding performed. For the full system evaluation the MPE models were used and an additional adaptation/lattice-rescoring third pass (P3) performed. The P3 adaptation consists of constrained Maximum Likelihood Linear Regression (MLLR) and lattice-based MLLR adaptation, again followed by CN decoding. For system combination the output from multiple P3-passes were combined using ROVER [16]. A more detailed system description can be found in [15].

For morpheme-to-word conversion the MARIE[4] N-gram SMT decoder was used. WERs are given in % and WERs for the MADA-based systems were obtained after morpheme-to-word conversion. In this work the mapping was based on a 3-gram LM built on the acoustic training data transcriptions ($\sim$11M words). After normalisation and MADA processing, the MADA-domain data was aligned with the original data. The resulting streams of source-target token pairs were used to train the LM which generated the mappings. The OOT rates were 1.6% for dev07, and dev08, and 3.8% for eval07. Applying the fall-back procedure for unseen OOT tokens described in Section 3, this was effectively lowered to 0.8% for dev07 and dev08, and 2.9% for eval07. The use of higher order 'bilingual' LMs for the morpheme-to-word conversion was also explored. Applying a 4-gram rather than a 3-gram LM did not give any improvements. Similarly, the training data used to train the bilingual mapping LMs was increased to $\sim$154M words. However, though the OOT rates were somewhat reduced, no improvement in system performance was observed.

To compare the word-based and morpheme-based systems, OOV rates for three STT Arabic development sets `dev07`, `eval07`,[5] and `dev08` were computed. Each of these testsets consists of approximately 20k Arabic words (corresponding to 2.5-3 hours of audio) selected from both broadcast news and broadcast conversation sources. These testsets are used by the various sites participating in the DARPA GALE program. The number of morphemes per word and the OOV rates are given in Table 2. Those OOV rates for the D2 and D2+Al systems have been normalised using the equation

$$\%\mathrm{OOV}_{\mathrm{norm}} = \%\mathrm{OOV} \times \frac{\#\text{ of morphemes}}{\#\text{ of words}}. \qquad (2)$$

The similar OOV rates for D2+Al compared to D2 in Table 2 are partly related to the OOV rate normalisation.

System	morph/word	dev07	eval07	dev08
Word	-	1.25	2.72	1.20
D2	1.19	0.79	0.54	0.63
D2+Al	1.36	0.76	0.54	0.63

Table 2: OOV rates and morpheme-to-word ratios.

4.2. Development Results

For system development the ML trained acoustic models together with the LMs trained on 8 sources and the P1-2 decoding was used. In Table 3 the word-based system is compared with the D2 system after morpheme-to-word conversion. Both 3-gram and 1-gram SMT mapping results are given. Note that in the 1-gram case the LM effectively reduces to a look-up table relating the most likely MADA-to-word token pairs. The D2 MADA-based system out-performed the word-based one by 0.5% up to 1.4% absolute. Comparing the 1-gram mapping and the 3-gram mapping, indicates that the context information used by the 3-gram SMT accounts for up to 0.2% of the absolute gains in WER. The D2 outcomes were also scored without morpheme-to-word conversion but after rejoining prefixes to their subsequent stems. This always resulted in a performance loss compared to the word based system. To investigate which aspect of the morphological analysis caused the observed performance gains, the decomposition, or the stem normalisation, a system (D2+rejoin) was built in which the affixes were rejoined to their respective stems after the D2 processing stage. This produced data which contained the same number of tokens as the word-based data, with the difference that the word stems have been subjected to MADA normalisation. Table 3 also gives these results. Comparing the D2 with the D2+rejoin results show that, though not consistent over-all test sets, there are gains from stem normalisation, up-to 0.4% absolute. Thus the gains from the MADA-based system are split between the decomposition and the stem normalisation.

System		dev07	eval07	dev08
Word	-	18.8	20.0	21.8
D2	1-gram SMT	18.2	18.8	21.3
	3-gram SMT	18.2	18.6	21.2
D2+rejoin	3-gram SMT	18.5	19.6	21.8

Table 3: % WER for the word-based system compared to a D2 system after morpheme-to-word back-mapping and solely morpheme rejoining.

Next, the D2-Al configuration was investigated. In this case, as the D2-Al configuration exhibits a higher morpheme-to-word ratio than the D2 configuration (see Table 2), the impact on the LM span must be considered. Table 4 provides the corresponding results for applying a 3-gram, a 4-gram, and a 5-gram LM during the lattice rescoring stage.

[4] Available from http://gps-tsc.upc.es/veu/soft/soft/marie/.

[5] The non-sequestered version of the GALE `eval07` testset was used.

System	LM-order	dev07	eval07	dev08
Word	**3-gram**	19.1	20.3	22.0
	4-gram	18.8	20.0	21.8
	5-gram	18.7	20.1	21.7
D2	**3-gram**	19.2	19.4	21.9
	4-gram	18.2	18.6	21.2
	5-gram	18.1	18.6	21.2
D2+Al	**3-gram**	22.3	22.5	24.9
	4-gram	18.6	19.4	21.7
	5-gram	18.5	19.4	21.6

Table 4: The impact of LM span on the performance of the word-based, D2, and D2+Al systems (%WER).

For both MADA-based systems increasing the LM span from 3 to 4 resulted in large reductions in WER. The gains for the word-based system were smaller. For all systems, increasing the LM span to 5-grams gave only minimal improvements. The D2 system consistently outperformed the word-based and the D2+Al systems. However it is worth emphasising that a 3-gram LM was used in the initial lattice generation stage. This may have a slightly larger impact on the D2+Al system than the D2 system.

4.3. Full System Results

From the above results, the D2 system was developed further. D2-MADA-based MPE trained graphemic (G_{D2}) acoustic models and 24-source LMs were built. For contrast a word-based system was also built (G_{word}). Since large gains have previously been observed by combining graphemic and phonetic systems, a word-based vowelised (V_{word}) system was used in system combination (as described in Section 4.1). The P3 decoding setup (see Section 4.1) including acoustic model adaptation was used, followed by ROVER combination.

System		dev07	eval07	dev08
P3a	G_{word}	13.1	14.4	15.2
P3b	G_{D2}	12.5	13.6	14.2
P3c	V_{word}	11.6	13.2	14.2
ROVER	**P3a** $\oplus$ **P3c**	11.5	12.7	13.4
	P3a $\oplus$ **P3b**	12.2	13.2	13.8
	P3b $\oplus$ **P3c**	11.0	12.1	13.0

Table 5: MPE P3b decoding and ROVER results after MADA-to-word back-mapping, 24-source LM, WER in %. The differences between the lowest WER systems and the relevant baseline systems were statistically significant (using NIST's sctk-1.2 tests).

The P3 results show that the D2 gains are preserved for this more complicated adaptation configuration. The G_{D2} system outperformed the G_{word} system by 0.6%-1.0% absolute WER. However, there remained a performance gap to the V_{word} system.

Further improvements were obtained by ROVER combination. The graphemic word and D2 combination P3a $\oplus$ P3b outperformed the best graphemic system, G_{D2}, by 0.3%-0.4% in absolute WER. The phonetic-MADA-graphemic P3b $\oplus$ P3c combination outperformed the phonetic-graphemic baseline combination P3a $\oplus$ P3c by 0.4%-0.6% absolute WER (3.0%-4.7% relative). MADA processing was only applied to the graphemic systems. Further gains might come from extending the proposed morphological analysis framework to the vowelised system.

5. Conclusions

This work has investigated applying MADA morphological analysis and decomposition for Arabic broadcast news and conversation transcription. The MADA tools performs both morphologically decomposition and the stem normalisation. As word-based scoring, and possibly combination, is required, a novel morpheme-to-word conversion method was introduced to deal with the stem normalisation. This method, based on an N-gram Statistical Machine Translation approach, facilitates the back-mapping of the morpheme-based recognition hypothesis to the word-domain. The MADA D2 configuration was found to yield the best performance. Both the decomposition and stem normalisation schemes were found to be important for best recognition performance. Evaluating the MADA-decomposed system in an multi-pass adaptation/combination framework, showed gains of 0.6%-1.0% absolute WER over the word-based baseline.

6. References

[1] K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh & A. Stolcke, "Morphology-based language modeling for conversational Arabic speech recognition", Computer Speech & Language, Vol. 20(4), pp. 589-608, 2006.

[2] M. Afify, R. Sarikaya, H.-K.J. Kuo, L. Besacier & Y. Gao, "On the use of morphological analysis for dialectal Arabic speech recognition", Proc. ICSLP'06.

[3] D. Vergyri, K. Kirchhoff, K. Duh & A. Stolcke, "Morphology based language modeling for Arabic speech recognition", Proc. ICSLP'04.

[4] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz & J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription", Proc. ICASSP'06.

[5] G. Choueiter, D. Povey, S.F. Chen, & G. Zweig, "Morpheme-Based Language Modeling for Arabic LVCSR", Proc. ICASSP'06.

[6] L. Lamel, A. Messaoudi, & J.-L. Gauvain, "Investigating Morphological Decomposition for Transcription of Arabic Broadcast News and Broadcast Conversation Data", Proc. InterSpeech'08.

[7] N. Habash & F. Sadat, "Arabic preprocessing schemes for statistical machine translation", Proc. HLT-NAACL'06.

[8] O. Bender, E. Matusov, S. Hahn, S. Hasan, S. Khadivi & H. Ney, "The RWTH Arabic-to-English spoken language translation system", Proc. ASRU'07.

[9] A. de Gispert, G. Blackwood, J.J. Brunning & W.J. Byrne, "The CUED NIST 2008 Arabic-English SMT system", Proc. NIST Open Machine Translation 2008 Evaluation Workshop.

[10] D. Vergyri, A. Mandal, W. Wang, A. Stolcke, J. Zheng, M. Graciarena, D. Rybach, C. Gollan, R. Schlüter K. Kirchhoff, A. Fari & N. Morgan, "Development of the SRI/Nightingale Arabic ASR system", Proc. Interspeech'08.

[11] F. Sadat & N. Habash, "Combination of Arabic preprocessing schemes for statistical machine translation", Proc. COLING/ACL'06,

[12] N. Habash & O. Rambow, "Arabic tokenisation, part-of-speech tagging and morphological disambiguation in one fell swoop", Proc. ACL'05.

[13] J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa & M.R. Costa-Jussa, "N-gram-based Machine Translation", Computational Linguistics, Vol. 32, No. 4, pp. 527–549, 2006.

[14] J.M. Crego, J.B. Mariño & A. de Gispert, "An N-gram-based Statistical Machine Translation Decoder", Proc. Interspeech'05.

[15] M.J.F. Gales, F. Diehl, C. K. Raut, M. Tomalin, P.C. Woodland & K. Yu, "Development of a phonetic system for large vocabulary Arabic speech recognition", Proc. ASRU'07.

[16] J.G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)", Proc. ASRU'97.