The Effects of Factorizing Root and Pattern Mapping in Translating between Tunisian Arabic and Standard Arabic

Ahmed Hamdi[1] **Rahma Boujelbane**[1,2] **Nizar Habash**[3] **Alexis Nasr**[1]
(1) Laboratoire d'Informatique Fondamentale de Marseille, Aix-Marseille University
(2) Multimedia, InfoRmation Systems and Advanced Computing Laboratory
(3) Center for Computational Learning Systems Columbia University
{ahmed.hamdi,rahma.boujelbane,alexis.nasr}@lif.univmrs.fr
habash@ccls.columbia.edu

Abstract

The development of natural language processing tools for dialects faces the severe problem of lack of resources. In cases of diglossia, as in Arabic, one variant, Modern Standard Arabic (MSA), has many resources that can be used to build natural language processing tools. Whereas other variants, Arabic dialects, are resource poor. Taking advantage of the closeness of MSA and its dialects, one way to solve the problem of limited resources, consists in performing a translation of the dialect into MSA in order to use the tools developed for MSA. We describe in this paper an architecture for such a translation and we evaluate it on Tunisian Arabic verbs. Our approach relies on modeling the translation process over the deep morphological representations of roots and patterns, commonly used to model Semitic morphology. We compare different techniques for how to perform the cross-lingual mapping. Our evaluation demonstrates that the use of a decent coverage root+pattern lexicon of Tunisian and MSA with a backoff that assumes independence of mapping roots and patterns is optimal in reducing overall ambiguity and increasing recall.

1 Introduction

The Arabic language has many variants. Modern Standard Arabic (MSA) is one of them. It is the official language of all Arab countries. However, MSA is the native language of no Arabic speakers. It is used for education, printed and spoken media. There exists also a variety of Arabic dialects which are the native languages of Arabic speakers. Unlike MSA, Dialectal Arabic (DA) varieties are only spoken. Therefore, there is no standard orthographic conventions (Habash, 2010; Habash et al., 2012b).

Most of the Arabic natural language processing (NLP) resources are built in order to process MSA. Very few works on processing dialects have been established, and mainly for Egyptian, Iraqi and Levantine Arabic. In this work, we focus on the Tunisian Arabic dialect (TUN), an important yet less studied Arabic dialect. We propose to transform it into a form that is close to MSA by using morphological analysis and generation in order to take advantage of MSA NLP tools. Our approach relies on modeling the translation process over the deep morphological representations of roots and patterns, commonly used to model Semitic morphology. We compare different techniques for how to perform the cross-lingual mapping. Our evaluation demonstrates that the use of a decent coverage root+pattern lexicon of Tunisian and MSA with a backoff that assumes independence of mapping roots and patterns is optimal in reducing overall ambiguity and increasing recall.

The paper is organized as follows. We first present some related work in the next section. Section 3 discusses similarities and differences between MSA and TUN verbal morphology. In Section 4, we describe different tools that are used throughout this work. Section 5 evaluates our system.

2 Related Work

A limited amount of work has been done on building DA resources and tools, and mainly for Egyptian, Iraqi and Levantine Arabic. Maamouri et al. (2004b) presented a transcription corpus with its

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 125–134.

design principles, development tools and guidelines for speech recognition research. Habash et al. (2012b) developed a conventional orthography for dialectal Arabic (CODA) designed for developing computational models of Arabic dialects. CODA was used in the design of a morphological analyzer for Egyptian Arabic (Habash et al., 2012a), as well as a morphological disambiguation system for Egyptian Arabic (Habash et al., 2013) and a system for normalizing spontaneous orthography (Eskander et al., 2013). A morphological analyzer and generator for Arabic dialects (MAGEAD) was also developed for MSA and Levantine Arabic (Habash et al., 2005; Habash and Rambow, 2006; Altantawy et al., 2010; Altantawy et al., 2011). Al-Sabbagh and Girju (2010) described an approach of mining the web to build a DA-to-MSA lexicon. Riesa and Yarowsky (2006) presented a supervised algorithm for online morpheme segmentation on DA that cut machine translation out-of-vocabulary (OOV) words by half. Zbib et al. (2012) demonstrated an approach to cheaply obtaining DA-English data using crowd-sourcing.

Several researchers have considered the idea of exploiting existing MSA rich resources to build tools for DA NLP. For example, in order to use MSA treebanks to parse Levantine Arabic, Chiang et al. (2006) compared three methods that rely on translating between MSA and Levantine. Abo Bakr et al. (2008) introduced a hybrid approach to transfer a sentence from Egyptian Arabic into MSA. Sawaf (2010), Salloum and Habash (2011) and Salloum and Habash (2013) converted DA into MSA using a dialectal morphological analyzer and various mapping rules. Salloum and Habash (2011)'s DA morphological analyzer (ADAM), was built by extending a MSA analyzer in a noisy fashion. Their goal was to maximize analyzability not correctness. Mohamed et al. (2012) described a method for translating disambiguated MSA to Egyptian Arabic using a rule-based system. Their system reduced OOVs and improved POS tagging accuracy.

In this paper, we explore a similar approach to previous efforts (Sawaf, 2010; Mohamed et al., 2012; Salloum and Habash, 2013) but using a well-motivated deep morphological representation based on the MAGEAD approach (Habash and Rambow, 2006). Our solution is bi-directional unlike previous efforts and we demonstrate our approach on Tunisian Arabic.

3 Morphology: MSA vs Tunisian Arabic

Many similarities and differences exist between MSA and TUN in every aspect of verbal morphology: cliticization, inflection and derivation.

3.1 Cliticization Morphology

Various particles, called clitics, attach to inflected words. Clitics are optional and do not change the core meaning of the verbs they attach to. There are two main differences in cliticization morphology between MSA and Tunisian. First, several MSA clitics change their form in Tunisian. For example, the MSA interrogative particle proclitic (prefixing clitic) +أ *Âa+*[1] becomes the enclitic (suffixing clitic) ش+ *+š*. Second, some MSA clitics become detached in TUN and vice versa. The MSA future particle proclitic +سـ *sa+* is realized as the autonomous particle باش *bAš* with TUN verbs. Inversely, indirect object pronouns are realized as enclitics in TUN verbs and not in MSA. The general structure of MSA and TUN verbs is represented in the following two regular expressions:

QST? CNJ? PRT? MSA_VERB PRN_D?

CNJ? PRT? TUN_VERB PRN_D? PRN_I? (NEG|QST)?

QST (question) is the interrogative particle, CNJ is either the conjunctions +و *w+* 'and' or +ف *f+*'so'. PRT is the class of particle proclitics such as future, prepositional and negation particle. NEG is a negation enclitic specified for TUN used with a negation proclitic. PRN_D and PRN_I are the direct and indirect object pronouns, respectively.

3.2 Inflectional and Derivational Morphology

Arabic words are constructed using two kinds of morphological operations: templatic and affixational. Functionally, both operations are used inflectionally or derivationally (Habash, 2007). In templatic morphology, a typically triliteral root and a pattern combine to form a word's stem, which is then extended with prefixes and suffixes, e.g., the TUN verb ومانقارنوهاش *wmAnqArnuwhAš* 'and we do not compare her/it' can be analyzed as *w+mA+n-*{$\frac{1A23}{\sqrt{qrn}}$}*-uw+hA+š*, where 1A23 is the

[1]Arabic orthographic transliteration is presented in the HSB scheme (Habash et al., 2007): (in alphabetical order)
ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
A b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional letters: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ىء , ħ ة, ý ى.

pattern, $\sqrt{qrn}$ the root, clitics are marked with '+' delimiter and affixes with '-' delimiter. MSA has a richer inflectional morphology than TUN. In fact, some MSA features such as nominal case and verbal mood do not exist in TUN. Furthermore, the MSA number values of singular, dual and plural are reduced to singular and plural. Masculine and feminine values of gender feature are not distinguished in TUN except for the third person singular. Patterns carry a general meaning, the MSA pattern Ai12a33, for example, denotes the change of state. This pattern is not used in TUN and Tunisians express the state change by using the pattern 12A3 which not exists in MSA. Furthermore, some MSA patterns are not defined in TUN and vice versa.

4 Tools and Resources

Our architecture relies on the morphological processing tool MAGEAD and on a transfer lexicon.

4.1 MAGEAD

MAGEAD (Habash and Rambow, 2005) is a morphological analyzer and generator for the Arabic language family (MSA and Arabic dialects). MAGEAD relates (bidirectionally) a lexeme and a set of linguistic features to a surface word form through a sequence of transformations. In a generation perspective, the features are translated to abstract morphemes which are then ordered, and expressed as concrete morphemes. The concrete templatic morphemes are interdigitated and affixes added, finally morphological and phonological rewrite rules are applied.

4.1.1 Lexeme and Features

Morphological analyses are represented in terms of a lexeme and features. The lexeme is defined as a root, a morphological behavior class (MBC). We use as our example the surface form ازدهرت *Aizdaharat* 'she flourished'. The MAGEAD lexeme-and-features representation of this word form is as follows:

(1) Root:zhr MBC:verb-VIII POS:V PER:3 GEN:F NUM:SG ASPECT:PERF

4.1.2 Morphological Behavior Class

An MBC maps sets of linguistic feature-value pairs to sets of abstract morphemes. For example, MBC verb-VIII maps the feature-value pair ASPECT:PERF to the abstract root morpheme [PAT_PV:VIII], which in MSA corresponds to the concrete root morpheme V1tV2V3, while the MBC verb-II maps ASPECT:PERF to the abstract root morpheme [PAT_PV:II], which in MSA corresponds to the concrete root morpheme 1V22V3. MBCs are defined using a hierarchical representation with non-monotonic inheritance. The hierarchy allows to specify only once those feature-to-morpheme mappings for all MBCs which share them. For example, the root node of MSA MBC hierarchy is a word, and all Arabic words share certain mappings, such as that from the linguistic feature conj:w to the clitic w+. This means that all Arabic words can take a cliticized conjunction. Similarly, the object pronominal clitics are the same for all transitive verbs, no matter what their templatic pattern is.

4.1.3 MAGEAD Morphemes

To keep the MBC hierarchy variant-independent, a variant-independent representation of the abstract morphemes (AMs) that the MBC hierarchy maps to have been chosen. The AMs are then ordered into the surface order of the corresponding concrete morphemes. The ordering of AMs is specified in a variant-independent context-free grammar. At this point, our example (1) looks like this:

(2) [Root:zhr][PAT_PV:VIII][VOC_PV:VIII-act] + [SUBJSUF_PV:3FS]

Note that the root, pattern, and vocalism are not ordered with respect to each other, they are simply juxtaposed. The '+' sign indicates the ordering of affixational morphemes. Only now are the AMs translated to concrete morphemes (CMs), which are concatenated in the specified order. Our example becomes:

(3) <zhr,V1tV2V3,iaa> + at

Simple interdigitation of root, pattern and vocalism then yields the form iztahar+at.

4.1.4 MAGEAD Rules

MAGEAD uses two types of rules. Morphophonemic/phonological rules map from the morphemic representation to the phonological and orthographic representations. Orthographic rules rewrite only the orthographic representation. For our example, we get /izdaharat/ at the phonological level (as opposed to /iztaharat/). Using standard MSA diacritized orthography, our example becomes *Aizdaharat*. Removing the diacritics turns this into the more familiar *Azdhrt*. We follow

(Kiraz, 2000) in using a multi-tape representation. MAGEAD extend the analysis of Kiraz by introducing a fifth tier. The five tiers are used as follows: Tier 1: pattern and affixational morphemes; Tier 2: root; Tier 3: vocalism; Tier 4: phonological representation; Tier 5: orthographic representation. In the generation direction, tiers 1 through 3 are always input tiers. Tier 4 is first an output tier, and subsequently an input tier. Tier 5 is always an output tier.

4.1.5 From MSA to Tunisian

We adapted MAGEAD to process TUN verbs. Our effort concentrated on the orthographic representation. Changes concerned only the representation of linguistic knowledge, leaving the processing engine unchanged. We modified the MBC hierarchy, adding one MBC, removing three and editing five. The AM ordering has been modified and a new AM has been added for indirect object. The mapping from AMs to CMs and the definition of rules, which are variant-specific, are obtained from a linguistically trained native speaker. Furthermore, we needed to change some morphophonemic rules. In MSA, for example, the gemination[2] rule, allows deleting the vowel between the second and the third radical if it is followed by a suffix starting with a vowel: compare مددت *madad+tu* 'I extended' with مدّت *mad∼+at* 'she extended' (NOT *madad+at*). In Tunisian, in contrast, gemination always happens, independently of the suffix: مدّيت *mad∼+iyt* 'I extended' and مدّت *mad∼+it* 'she extended'. Many other rule changes were needed for TUN. For example, the first root radical becomes a long vowel in the imperfective aspect when it corresponds to ء *'* (*hamza/glottal stop*) (يأكل *yÂkl* becomes ياكل *yAkl* 'he/it eats'). On the other hand, verbs whose root ends with ء *'*, behave the same way as verbs whose final root radical ي *y* in the perfective aspect. For example, roots of TUN verbs بدينا *bdynA* 'we started' and رمينا *rmynA* 'we threw' are respectively ب د ء *b d '* and ر م ي *r m y*. More details are discussed in Hamdi et al. (2013).

4.2 Root and Pattern Lexicon

Our lexicon is made of pairs of the form (P_{MSA}, P_{TUN}) where P_{MSA} and P_{TUN} are themselves pairs made of a root and an MBC. Its development was based on the Arabic Tree Bank (ATB) (Maamouri et al., 2004a) which contains 29,911 verb tokens. In order to extract the lemmas and the roots of these verbs, we used the morphological analyzer ElixirFM (Smrž, 2007) which extracts the lemma and the root of MSA inflected forms.[3] Then, each token of MSA lemma was translated by a Tunisian native speaker. At this point, lexicon entries are composed of a lemma and a root on the MSA side but only a lemma on the TUN side. We then associated to every entry an MBC (on the MSA side) and an MBC and a root (on the TUN side). In 81.49% of cases, we identified an Arabic existing root for TUN verbs. When there was no root for a given lemma, we used a deductive method to create a new one. Indeed, given the equation root + pattern = lemma, when we have a lemma and a pattern, it is possible to deduce a root. Using this process, we defined 100 new specific Tunisian roots.

In its current state, the lexicon contains 1,638 entries. The TUN side contains 920 distinct pairs and the MSA side 1,478 distinct pairs. As expected, the ambiguity is more important in the TUN $\rightarrow$ MSA sense. On average, a TUN pair corresponds to 1.78 MSA pairs, 1.11 in the opposite direction. The maximum ambiguity is equal to four in the MSA $\rightarrow$ TUN direction and sixteen in the opposite direction. More will be said about ambiguity in Section 5.

A sample of the lexicon appears in Table 1. The MBC indicates the pattern and in some cases the short vowels of the second root radical in the perfective and the imperfective aspects since they could change from verb to other. As shown in the table, a MSA MBC could be mapped to many TUN MBCs and vice versa.

Two by-products can be built form the lexicon, a root lexicon and a pattern correspondence table, both described below.

4.2.1 Root Lexicon

The root lexicon is made of pairs of the form (r_{MSA}, r_{TUN}), where r_{MSA} is an MSA root and r_{TUN} is a TUN root. The root lexicon contains 1,329 entries. The MSA side contains 1,050 dis-

[2]The second and the third root radical are identical.

[3]We did not use MAGEAD to perform the root extraction because the work on the lexicon had already started independently. MAGEAD for MSA, whose lexicon is based on the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002) – just like ElixirFM, could have been used in principle.

MSA		TUN		English
Root	**MBC / Pattern**	**Root**	**MBC / Pattern**	**Gloss**
Smt	1-aa / 1a2a3	skt	1-ii / 12i3	'to be silent'
Hlq	1-aa / 1a2a3	Hjm	2-ii / 1a22i3	'to cut hair'
rtb	2 / 1a22a3	nZm	2-ii / 1a22i3	'to rank'
Hlq	2 / 1a22a3	Tyr	1-a / 12a3	'to fly'
xSm	3 / 1A2a3	ʕrk	3-ii / 1A2i3	'to dispute'
dhm	3 / 1A2a3	hjm	1-ii / 12i3	'to attack'
bhr	4 / Aa12a3	ʕjb	1-ii / 12i3	'to amaze'
xfy	4 / Aa12a3	xby	2-ai / 1a22a3	'to hide'
ršf	5 / ta1a22a3	ršf	5-ii / t1a22i3	'to savor'
ʕjb	5 / ta1a22a3	bht	1-ii / 12i3	'to be surprised'
šjr	6 / ta1A2a3	ʕrk	6 / t1A2i3	'to fight'
ʕfy	6 / ta1A2a3	bry	1-aa / 12a3	'to be cured'
xfD	7 / Ain1a2a3	nqS	1-uu / 12u3	'to decrease'
sHb	7 / Ain1a2a3	bTl	2-ii / 1a22i3	'to step down'
nhy	8 / Ai1ta2a3	kml	1-ii / 12i3	'to be end'
Hdn	8 / Ai1ta2a3	Hml	2-ii / 1a22i3	'to hold'
dʕy	10 / Aista12a3	ʕdy	10 / Aista12a3	'to invite'
wfy	10 / Aista12a3	kml	2-ii / 1a22i3	'to complete'

Table 1: A sample TUN-MSA lexicon. The pattern provided is the form used with 3rd masculine singular perfective inflection. It is only presented for illustrative reasons to exemplify and highlight differences between TUN and MSA MBCs.

tinct roots and the TUN side 646 ones. 519 entries are composed of the same root on both sides. As in the root and pattern lexicon, the ambiguity is higher in the TUN → MSA direction. On average, a TUN root is paired with 2.06 MSA roots. In the opposite direction, this figure is equal to 1.26.

4.2.2 Pattern Correspondence Table

The pattern correspondence table indicates, for a pattern in MSA or TUN, the most frequent corresponding pattern in the other side. The pattern correspondence table is itself built on a pattern correspondence matrix, which is represented in Table 2. Each line of the matrix corresponds to a MSA pattern and each column to a TUN pattern. The matrix reads as follow, MSA pattern 1, for example, corresponds in 434 times to TUN pattern 1, 98 times to TUN pattern 2, and so on.

This matrix reveals several interesting facts. First, all patterns are not present in MSA or TUN in our lexicon. Pattern 9, for example is absent both in MSA and TUN and patterns 4 and 7 are absent on the TUN side. Second, there is a general tendency to keep the same pattern on the source and target sides of a lexicon entry. This is represented in the matrix by the fact that figures on the diagonal (in bold face) usually are the highest figure of both their line and column (the only exception is pattern 8). When a pattern does not exist in

		TUN						
		1	2	3	5	6	8	10
MSA	1	**434**$^{*}_{*}$	98	10	15		2	
	2	39	**298**$^{*}_{*}$	2	2	2		2
	3	24	19	**56**$^{*}_{*}$		2		
	4	69	**118***	4	6			
	5	26	16	2	**88**$^{*}_{*}$			3
	6	18	14	2	7	**26**$^{*}_{*}$		
	7	**13***	7	2				
	8	**41***	24	5	16	4	**18**$_{*}$	
	10	17	24	2	3			**31**$^{*}_{*}$

Table 2: Pattern correspondence matrix. Bolded cells are either the highest counts when translating from TUN to MSA or from MSA to TUN. X^{*} indicates highest count from MSA to TUN; and X_{*} indicates highest count from TUN to MSA.

TUN, it is usually mapped to pattern 1.

The extraction of the pattern correspondence tables form the pattern correspondence matrix is straightforward: it consists in selecting for every pattern in the source side the most frequent pattern for the target side. It is interesting to note that is some cases, the most frequent pattern clearly dominates the other patterns, as it is the case for pattern 2 in MSA. In other cases, the tendency is not clear, as in pattern 4 in MSA.

Overall, the matrix tells us that selecting a target root and a target pattern are not independent processes. In other words, the root and pattern lexicon contains more information than the root lexicon along with the pattern correspondence table. We will experimentally quantify, in Section 5, the influence of making such an independence hypothesis.

5 Evaluation

The process of translating a source verbal form to a target verbal form proceeds in three main steps: morphological analysis using MAGEAD for the source language, followed by lexical transfer of roots and MBCs and finally, morphological generation of target verbal forms. All of these steps are reversible.

The whole process contains two sources of ambiguity: the analysis can create multiple (root, MBC) pairs and the lexicon may propose for an input pair many target pairs.

As we mentioned in the introduction, the goal of this work is not translation for TUN to MSA but

generating from a TUN text an approximation of MSA, so that MSA NLP tools, such as morpho-syntactic taggers or parsers can be applied to this new form of text with acceptable results. The experiments described here provide only a partial evaluation, they allow to measure the proportion of cases in which the correct MSA form is generated given a TUN form.

The evaluation process is faced with the problem of lack of written resources for dialects. To overcome this problem, we used a book by Dhouib (2007) which is a Tunisian theater piece. 1500 tokens of TUN verbal forms were identified and translated in context to MSA by two Tunisian native speakers. At the end of this process, 1500 pairs were produced. This set was divided into two equal parts. The first was used as a development set and the second as a test set. Two standard metrics were used to evaluate the process: recall, which indicates the proportion of cases where the correct target form was produced; and ambiguity, which indicates the number of target forms produced on average for an input. The development set allowed us to fill some gaps in MAGEAD and enrich our lexicon.

We conducted the evaluation on undiacritized verbal forms since most of written Arabic is undiacritized. Without neither morphological nor lexical transfer, recall reaches 30.93% on tokens and 29.44% on types[4] but ambiguity is still at 1.0. This experiment gives the ratio of identical undiacritized TUN and MSA verbal forms in the test set.

In the following four sections, we present a series of experiments with different ways of realizing the transfer especially with respect to factorizing roots and patterns.

5.1 Pattern Correspondence Table

The most simple transfer process that we have experimented consists in leaving the source root unchanged and selecting the target pattern by a pattern correspondence table lookup. This experiment corresponds to the situation in which we do not have at our disposal a transfer lexicon. Since pattern is defined as a superset of MBCs, the target pattern maps to many target MBCs, each of them is associated to the target root and features to form the input of the morphological generator. We have chosen to build a correspondence pattern table instead of a correspondence MBC table for two main reasons : first, evaluations are made in an undiacritized set of verbs. Second, patterns carry a general meaning which can be a way to match MSA with TUN patterns. A block diagram of the process is presented in Figure 1 and the result of the experiment can be found in Table 3.

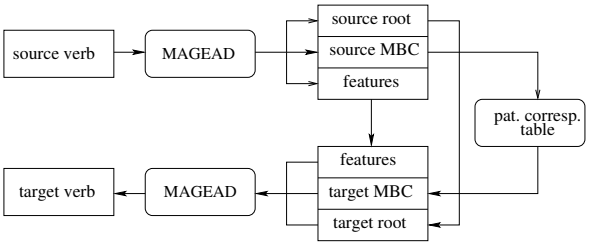

Figure 1: Translation process of source verbal form to target verbal form using a pattern correspondence table

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	47.74	43.40	39.41	37.61
MSA → TUN	52.55	48.05	5.89	7.12

Table 3: Recall and ambiguity on test set using pattern correspondence table

Table 3 shows two interesting features. First, the recall is quite low, around 50%. Keeping the source root is therefore a very rough approximation of the target variant. Second, the ambiguity is much higher in the TUN→MSA direction. This is due to the fact that TUN forms are morphologically more ambiguous than MSA forms. On average, a TUN form has 24.05 different analyses while MSA forms has on average 10.21 analyses. As mentioned in Section 3 MSA has a richer inflectional morphology than TUN, however our system used the same features for TUN and MSA analysis. Consequently, when a feature does not exist on TUN side, it produces many identical analysis with different values of this feature and generates subsequently many MSA verbal forms.

The same experiment was done using two target patterns instead of one (see Table 4). Table 4 shows a slight increase in recall. It rises on tokens to 51.65% in the TUN→MSA direction and 53.96% in the other direction. However, the ambiguity becomes higher, the process produces about

[4]Types are unique instances of tokens.

70 MSA verbs on average for a TUN token.

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	51.65	48.23	66.98	64.69
MSA → TUN	53.96	50.87	9.81	10.68

Table 4: Recall and ambiguity on test set using pattern correspondence table

5.2 Root Lexicon and Pattern Correspondence Table

In this experiment, the target pattern is selected as before by a lookup in the pattern correspondence table but the target roots are selected by a root lexicon lookup. This new setting was devised in order to increase the recall by better modeling root modification. The block diagram of the new setting appears in Figure 2 and the results on test set in Table 5 and 6.

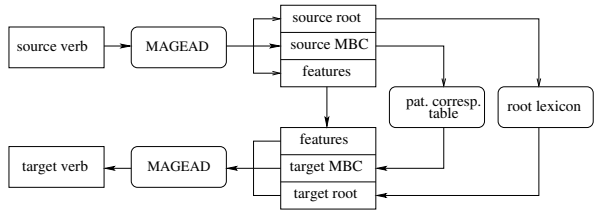

Figure 2: Translation process of source verbal form to target verbal form using a root lexicon and a pattern correspondence table

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	68.98	66.56	74.37	72.89
MSA → TUN	72.37	71.60	13.70	14.52

Table 5: Recall and ambiguity on test using a root lexicon and a pattern correspondence table

As expected, Table 5 shows a significant improvement of the recall. Ambiguity has also increased, this is due to the fact that a source root can map to several target roots: on average 2.06 in the TUN→MSA direction and 1.26 in the opposite direction.

Using the two most frequent target patterns from the pattern correspondence table, the translation process gives the highest recall and ambiguity, as shown in Table 6. In the MSA→TUN direction, recall rises to 86.12% on tokens and 81.77% in the inverse direction. The downside of this process is the ambiguity which becomes more then 100 in the TUN→MSA direction.

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	81.77	80.66	126.44	122.45
MSA → TUN	86.12	84.97	21.92	22.56

Table 6: Recall and ambiguity on test using a root lexicon and a pattern correspondence table

5.3 Root and Pattern Lexicon

In the preceding experiment, target roots and target patterns are translated independently and paired to compose the input of the morphological generator. But, as mentioned in Section 1, target root selection and target pattern selection are not independent processes: two source (root, pattern) pairs, sharing a common pattern can select different target patterns. In such cases the preceding method will give birth to incorrect (root, pattern) pairs and, eventually, incorrect verbal forms. In this experiment, target roots and patterns are selected together by a root and pattern lexicon access. The new process is represented in Figure 3 and results appear in Table 7.

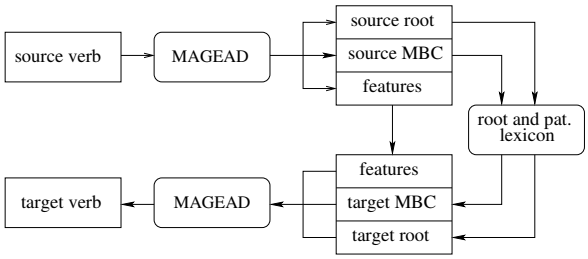

Figure 3: Translation process of source verbal form to target verbal form using a root and pattern lexicon

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	76.43	74.52	26.82	25.57
MSA → TUN	79.24	75.10	1.47	3.10

Table 7: Recall and ambiguity on test using a root and pattern lexicon

Replacing the root lexicon and the pattern correspondence table by a root and pattern lexicon has

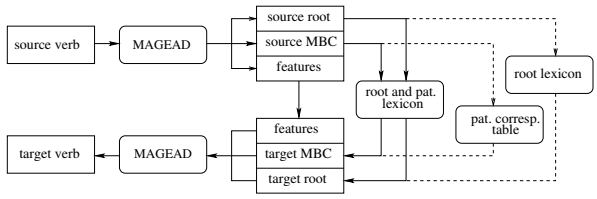

Figure 4: Translation process of source verbal form to target verbal form using a root and pattern lexicon with backoff on a root lexicon and a pattern correspondence table

a positive effect both on recall and ambiguity. The difference between the results of this experiment and the preceding one allows us to quantify the independence hypothesis of the root selection and the pattern selection we made in the preceding experiment.

The main weakness of this method is lexical coverage. We cannot expect to have a complete root and pattern lexicon and, sometimes, lexicon access fails. It is interesting at this point to mention the results of the same experiment on the development set. Recall that the verbal forms included in the development set have been used to populate the lexicon. As a consequence, a lexicon access never fails, and always produces the correct target (root, pattern) pair. The results of such an experiment, although artificial, allow to estimate an upper bound of such a method. In TUN $\rightarrow$ MSA direction, recall on tokens reaches 87.65% and in the inverse direction, it reaches 89.56%.

The reason why we did not reach 100% recall in this experiment is due to the fact that both MSA and TUN MAGEAD do not always produce the correct analysis, when used as an analyzer, or the correct form when used as a generator. An error analysis in the TUN $\rightarrow$ MSA direction showed that 21.8% of errors come from MSA MAGEAD and 78.2% from TUN MAGEAD. Most MAGEAD mistakes are due to morphological phenomena which have not been implemented yet, as quadriliteral verbs and the imperative form of defective verbs.[5]

5.4 Root and Pattern Lexicon with Backoff

In order to deal with low lexical coverage, we devised a variant of the preceding method which backs off, in cases of lexicon lookup failure, to the root lexicon and a the pattern correspondence table. The architecture of the system is shown in Figure 4, where the dotted lines represent the backoff path.

As Table 8 shows, this method increases recall significantly. This increase is itself the result of a better coverage. Ambiguity has also increased, this is due to the fact that when backing off, the transfer tends to be more ambiguous.

	recall		ambiguity	
	tokens	types	tokens	types
TUN $\rightarrow$ MSA	79.71	78.94	29.16	28.44
MSA $\rightarrow$ TUN	84.83	84.03	3.47	4.95

Table 8: Recall and ambiguity on test using a root and pattern lexicon with backoff on a root lexicon and a pattern correspondence table

6 Conclusion and Future Work

We presented a translation system between MSA and TUN verbal forms. This work is part of a wider project of translating Arabic dialects to an approximation of MSA. The results given by our system are about 80% recall in the TUN $\rightarrow$ MSA direction and 84% recall in the opposite direction. The translation process is highly ambiguous, in the MSA $\rightarrow$ TUN direction, the mean ambiguity is equal to 3.47 and reaches 29.16 in the opposite direction. A contextual disambiguation process is therefore necessary for such a process to be of practical use.

Future work will involve the development of a morphological model for nouns for TUN following the work of Altantawy et al. (2010), as well as a lexicon. In parallel we will work on the disambiguation of the TUN $\rightarrow$ MSA translations using a language model trained on a MSA corpus.

Acknowledgments

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Nos. HR0011-12-C-0014 and HR0011-12-C-0016. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

[5] Arabic defective verbs contain /w/ or /y/ in their root.

References

Abo Bakr, Hitham, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

Al-Sabbagh, Rania and Roxana Girju. 2010. Mining the Web for the Induction of a Dialectical Arabic Lexicon. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

Altantawy, Mohamed, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC. Valletta, Malta*.

Altantawy, Mohamed, Nizar Habash, and Owen Rambow. 2011. Fast Yet Rich Morphological Analysis. In *Proceedings of the 9th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2011)*, Blois, France.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2002L49.

Chiang, David, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of the European Chapter of ACL (EACL)*.

Dhouib, Elmoncef. 2007. *El Makki w-Zakiyya*. Publishing House Manshuwrat Manara, Tunis, Tunisia.

Eskander, Ramy, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Habash, Nizar and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Habash, Nizar and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia, July. Association for Computational Linguistics.

Habash, Nizar, Owen Rambow, and George Kiraz. 2005. Morphological Analysis and Generation for Arabic Dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24, Ann Arbor, Michigan.

Habash, Nizar, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In van den Bosch, A. and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., R. Eskander, and A. Hawwari. 2012a. A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

Habash, Nizar, Mona Diab, and Owen Rabmow. 2012b. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

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

Habash, Nizar. 2007. Arabic Morphological Representations for Machine Translation. In van den Bosch, A. and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, Nizar. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Hamdi, Ahmed, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. In *In proceedings of Traitement Automatique du Langage Naturel (TALN 2013)*.

Kiraz, George Anton. 2000. Multitiered nonlinear morphology using multitape finite automata: a case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105, March.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004a. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Maamouri, Mohamed, Tim Buckwalter, and Christopher Cieri. 2004b. Dialectal Arabic Telephone Speech Corpus: Principles, Tool design, and Transcription Conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools*.

Mohamed, Emad, Behrang Mohit, and Kemal Oflazer. 2012. Transforming standard arabic to colloquial arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–180, Jeju Island,

Korea, July. Association for Computational Linguistics.

Riesa, Jason and David Yarowsky. 2006. Minimally Supervised Morphological Segmentation with Applications to Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA06)*, pages 185–192, Cambridge,MA.

Salloum, Wael and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Salloum, Wael and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Sawaf, Hassan. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado.

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

Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada.

Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation

Sudip Kumar Naskar* **Antonio Toral** **Federico Gaspari** **Declan Groves**
Computer and System Sciences
Visva-Bharati University
India
sudip.naskar@visva-bharati.ac.in

School of Computing
Dublin City University
Ireland
{atoral, fgaspari, dgroves}@computing.dcu.ie

Abstract

Diagnostic evaluation of machine translation (MT) is an approach to evaluation that provides finer-grained information compared to state-of-the-art automatic metrics. This paper evaluates DELiC4MT, a diagnostic metric that assesses the performance of MT systems on user-defined linguistic phenomena. We present the results obtained using this diagnostic metric when evaluating three MT systems that translate from English to French, with a comparison against both human judgements and a set of representative automatic evaluation metrics. In addition, as the diagnostic metric relies on word alignments, the paper compares the margin of error in diagnostic evaluation when using automatic word alignments as opposed to gold standard manual alignments. We observed that this diagnostic metric is capable of accurately reflecting translation quality, can be used reliably with automatic word alignments and, in general, correlates well with automatic metrics and, more importantly, with human judgements.

1 Introduction

The study presented in this paper addresses the topic of diagnostic evaluation of machine translation (MT), which is receiving increasing attention due to its potentially crucial but still largely unexplored role in the development and subsequent deployment of MT systems. Diagnostic evaluation might be particularly useful to complement the overall system-level scores provided by automatic MT evaluation metrics. On the one hand, these automatic metrics represent cost-effective, objective and easily replicable measures, on the other, they provide only global indications that are normally too coarse to explain the performance of an MT system. An associated issue is that diagnostic evaluation needs to be as fine-grained as possible to be really useful in targeting specific weaknesses detected in MT output, for the system developers to be able to take corrective actions accordingly, and the users to asses the actual impact of the system's weaknesses.

This paper evaluates a diagnostic metric that assesses the performance of MT systems on user-defined linguistic phenomena. Focusing on English to French translation as a case study, the use of alternative automatic word alignments is investigated and compared against gold standard manual alignment to discuss how these different approaches impact on the results of diagnostic MT evaluation. The paper also presents a comparative evaluation of three MT systems judged according to standard automatic MT evaluation metrics, the diagnostic evaluation metric over a range of linguistic checkpoints, and human judgements. Additionally, we investigate how these different types of MT evaluation correlate to each other.

The paper is structured as follows. Section 2 presents previous work in diagnostic evaluation of MT, discussing the methodologies and tools that exist in this area, focusing in particular on the features of the diagnostic metric used in this study. Section 3 describes the datasets that were used for the experiments and Section 4 details the experi-

*Work done while at CNGL, School of Computing, Dublin City University.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 135–142.

mental setup. The results of the investigation are presented and analysed in Section 5, and finally some conclusions are drawn and possible avenues for future work are outlined in Section 6.

2 Related Work

Recognising that the ability to automatically identify and evaluate specific MT errors with diagnostic relevance is of paramount importance, Popović et al. (2006) propose a framework for the automatic classification of MT errors based on morpho-syntactic features. They show that linguistically-sensitive measures provide useful feedback to alleviate the problems encountered by MT. In a similar vein, Popović and Burchardt (2011) present a method for automatic error classification and compare its use with results obtained from human evaluation. They show good correlation between their automatic measures and human judgements across various error classes for different MT output.

Popović (2011) describes a tool for automatic classification of MT errors, which are grouped into five classes (morphological, lexical, reordering, omissions and unnecessary additions). The tool needs full-form reference translation(s) and hypotheses with their corresponding base forms. Additional information at the word level (such as PoS tags) can be used for a more delicate analysis. The tool computes the number of errors for each class at the document and sentence levels.

Max et al. (2010) propose an approach to contrastive diagnostic MT evaluation based on comparing the ability of different systems (or implementations of the same system) to correctly translate source-language words. Their contrastive lexical evaluation method does not rely on the direct comparison of the system's hypotheses with the reference translations, but for each source-language word it identifies which of the MT systems under consideration provide the correct output matching the reference. Their study is devoted to English–French and they point out the crucial role played by the quality of the alignment, suggesting that inaccuracies in the automatic alignment are bound to impair the reliability of this approach for lexical diagnostic evaluation.

Fishel et al. (2012) provide an overview of the field of diagnostic evaluation of MT, presenting a collection of freely available translation error-annotation corpora for various language pairs and comparing the performance of two state-of-the-art tools on automatic error analysis of MT.

Zhou et al. (2008) describe a tool for diagnostic MT evaluation called Woodpecker,[1] which is based on linguistic checkpoints. These are particularly interesting (or problematic) linguistic phenomena for MT processing identified by the user or developer who conducts the evaluation, e.g. ambiguous words, challenging collocations or PoS-n-gram constructs, etc. One needs to define a linguistic taxonomy which describes the phenomena to be captured in the diagnostic evaluation, deciding which elements of the source language one wants to investigate. This scheme is extremely flexible, and can be formulated at different levels of specificity, whereby the granularity of the checkpoints included depends on the objectives of the diagnostic evaluation.

While the notion of linguistic checkpoints is very useful within the context of diagnostic MT evaluation, Woodpecker has some limitations. First of all, language-dependent data for English–Chinese (the language pair covered in the study presented in (Zhou et al., 2008)) is hardcoded in the software, which therefore cannot be easily adapted to other language pairs. In addition, the licence with which Woodpecker is distributed (MSR-LA)[2] is quite restrictive, in that e.g. researchers cannot publicly release their own adaptations of the tool.

DELiC4MT[3](Toral et al., 2012) is a free open-source tool for diagnostic evaluation which offers similar functionality to Woodpecker. We chose to carry out experiments with DELiC4MT due to its language-independent nature. This recall-based diagnostic evaluation metric essentially works like other n-gram-based automatic MT evaluation metrics (i.e. counting n-gram matches between the MT output and the reference translations), except that it focuses on specific segments of the reference identified through linguistic constructs found in the source (i.e. linguistic checkpoints) and word alignment.

[1] http://research.microsoft.com/en-us/downloads/ad240799-a9a7-4a14-a556-d6a7c7919b4a/

[2] https://research.microsoft.com/en-us/projects/pex/msr-la.txt

[3] http://www.computing.dcu.ie/~atoral/delic4mt/

The final recall score produced by DELiC4MT is computed as in equation 1, where *R* is the set of references (*r*) of all the checkpoints (*c*) in *C*. A length-based penalty is introduced to penalise longer candidate translations (otherwise longer translations would have a better chance of returning higher scores) as in equation 2, where *length*(*C*) is the average candidate translation length and *length*(*R*) is the average reference translation length.

$$R(C) = \frac{\sum_{r \in R} \sum_{n-gram \in r} match(n-gram)}{\sum_{r \in R} \sum_{n-gram \in r} count(n-gram)} * penalty \quad (1)$$

$$penalty = \begin{cases} \frac{length(R)}{length(C)} & \text{if } length(C) > length(R) \\ 1 & otherwise \end{cases} \quad (2)$$

3 Datasets

The initial key decisions that had to be made to set up the experiment concerned the languages to be focused on as well as the domain and specific dataset to be selected for the investigation.

We decided to work on English–French, as human-aligned datasets are readily available for this language pair. We investigated a number of options in terms of manually annotated aligned English–French data to serve as gold standard, and considered, for example, using Biblical texts made available as part of the Blinker Annotation Project (Melamed, 1998). However, the syntax and vocabulary of this dataset presented some specific features which were not in line with actual uses envisaged for diagnostic evaluation in research or industrial settings.

The dataset that was chosen for our experiment was initially created for the shared task on word alignment held as part of the HLT/NAACL 2003 Workshop on Building and Using Parallel Texts (Mihalcea and Pedersen, 2003). The dataset used for this study consists of 447 English–French word-aligned sentence pairs drawn from the Canadian Hansard Corpus, consisting of parliamentary debates (Och and Ney, 2000), for a total of 7,020 tokens in English and 7,761 in French. It should be noted that we did not differentiate between 'sure' and 'probable' word alignments in this dataset and treat them as having the same weight.

Choosing a bilingual dataset from the domain of parliamentary speeches allowed us to conduct a fair and direct comparison with a closely related baseline English–French MT system built using the Europarl corpus[4] (Koehn, 2005).

4 Experimental Setup

4.1 MT Systems

We experimented with three MT systems: Google Translate[5], Systran[6] and a baseline Moses[7] system. Among the three MT systems, Google Translate and Moses are statistical MT systems while Systran is predominantly a rule-based system. The Moses system used for our experiments was trained on 3.6 million English–French sentence pairs taken from Europarl, the News Commentary corpus and a randomly selected section of the UN corpus. The system was tuned on a held-out development set consisting of 1,025 sentence pairs and used a 5-gram language model built using the SRILM toolkit (Stolcke, 2002).

4.2 Word Alignment

The diagnostic evaluation was carried out using both gold standard human alignments and three sets of automatic alignments. Thus, in total we carried out experiments on 4 different sets of word alignments. The idea behind this study was primarily to show whether the different possible alignments had an impact on the effectiveness of the diagnostic MT evaluation metric, also in comparison with gold-standard manual alignment and human evaluation.

We used GIZA++[8] (Och and Ney, 2003) to derive the automatic alignments between the source and target sides of the testset. We extracted three sets of alignments using the union, intersection and grow-diag-final heuristics, as implemented by the Moses training scripts. Since the testset is far too small to be accurately word-aligned using a statistical word-aligner and would suffer from data sparseness, additional parallel training data from the Europarl corpus was used. The additional training data was first tokenised, filtered (using source-target length ratio) and lower-cased. The testset was also subjected to tokenisation and lower-casing. The testset was then appended with

[4] http://www.statmt.org/europarl/
[5] http://translate.google.com
[6] http://www.systran.co.uk/
[7] http://www.statmt.org/moses/
[8] http://code.google.com/p/giza-pp/

the additional training data and word-aligned using GIZA++. Finally, from the word-alignment file only the word alignments for the sentences that correspond to the testset were extracted.

4.3 Linguistic Checkpoints

Regarding the linguistic phenomena, we considered a basic set of PoS-based checkpoints: adjectives (a), nouns (n), verbs (v), adverbs (r), determiners (dt), miscellaneous (misc), and pronouns (pro). The 'misc' checkpoint contains a variety of other PoS tags (CC, IN, RP and TO) (Santorini, 1990). We used Treetagger[9] (Schmid, 1994) to PoS-tag both sides of the testset.

It should be noted that the evaluation framework can potentially focus on more complex user-defined linguistic phenomena. In fact, it can be applied to a wide range of composite linguistic structures of interest to the MT developer or user for evaluation purposes. The metric can handle, e.g., combinations of literal words or lemmas with PoS tags. Evaluation on named entities and dependency structures is also supported by this diagnostic MT evaluation metric.

4.4 Human Judgements

In order to verify the results of the diagnostic evaluation, we carried out human evaluations on the output of the 3 different MT systems. These were done by 2 evaluators, both native French speakers and experienced in translation evaluation. They were asked to assign fluency and adequacy scores to the translations based on a discrete 5-point scale (LDC, 2005). In addition, they were asked to evaluate translation quality in terms of 5 PoS-based checkpoints (a, n, v, r and dt), again using a 5-point scale, with 1 representing instances where there were severe errors in the translation of all instances of the checkpoint and 5 indicating that all instances were translated perfectly. The evaluators were also asked to give a does-not-apply ('NA') score to sentences that did not contain the linguistic phenomenon under consideration.

5 Results

5.1 Diagnostic Evaluation

Table 1 shows the diagnostic evaluation results obtained on the gold standard word alignment. Tables 2, 3 and 4 present results obtained on the grow-diag-final, union and intersection alignments respectively. Each of these tables shows checkpoint-specific scores across systems. Table 1 shows in addition the number of checkpoint-specific instances (#Inst) extracted from the source side of the testset.

Checkpoint-specific statistically significant improvements are reported in these tables as superscripts. For representation purposes, we use *a*, *b*, and *c* for Google, Moses and Systran, respectively. For example, the Google score $0.4993^{b,c}$ for the adjective checkpoint in Table 1 means that the improvement provided by Google for this checkpoint is statistically significant over both Moses and Systran.

In addition to the checkpoint-specific scores, each of these tables provides an arithmetic mean (avg) and a weighted mean (w-avg, weighted by the number of instances for each checkpoint). The weighted average is considered as the system-level score for diagnostic evaluation. Tables 2, 3 and 4 also show the ratios with manual alignment (m-ratio). For example, Table 2 shows that the weighted means obtained by Google, Moses and Systran on grow-diag-final alignments are respectively 0.7337, 0.7132 and 0.7126 times those obtained on manual alignments.

	#Inst	Systems		
		Google	Moses	Systran
a	426	$0.4993^{b,c}$	0.4345	0.4369
n	1,649	$0.5420^{b,c}$	0.5025	0.5013
v	1,296	0.4037^{c}	0.3974^{c}	0.3603
r	348	0.4462	0.4198	0.4352
dt	824	$0.5968^{b,c}$	0.5479	0.5718
misc	1,079	$0.5788^{b,c}$	0.5376	0.5367
pro	428	$0.5740^{b,c}$	0.5049	0.5415
avg	6,050	0.5201	0.4778	0.4834
w-avg		0.5201	0.4831	0.4815

Table 1: Diagnostic evaluation results on manual alignments

As the scores in Table 1 suggest, Google clearly outperforms the other systems on all of the phenomena, and most of these improvements are statistically significant. The Moses baseline system performs slightly better than Systran according to the weighted averages. While some of the phenomena (e.g., nouns, verbs) are better handled by

[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

the Moses baseline system the scores in Table 1 also show that Systran performs quite better than this baseline system for adverbs, determiners and pronouns. This trend can be observed across Tables 2, 3 and 4 as well.

	Systems		
	Google	Moses	Systran
a	$0.3056^{b,c}$	0.2591	0.2440
n	$0.3374^{b,c}$	0.2958	0.2896
v	0.2583^{c}	0.2483^{c}	0.2272
r	0.3266^{b}	0.3061	0.3016
dt	$0.5117^{b,c}$	0.4621	0.4853
misc	$0.5199^{b,c}$	0.4698	0.4676
pro	0.4465^{b}	0.3976	0.4450
avg	0.3866	0.3484	0.3515
w-avg	0.3816	0.3445	0.3431
m-ratio	0.7337	0.7132	0.7126

Table 2: Diagnostic evaluation results on grow-diag-final alignments

	Systems		
	Google	Moses	Systran
a	$0.2748^{b,c}$	0.2281	0.2195
n	$0.3108^{b,c}$	0.2690	0.2650
v	0.2423^{c}	0.2305^{c}	0.2113
r	0.3191^{b}	0.3016	0.2937
dt	$0.4787^{b,c}$	0.4324	0.4552
misc	$0.4916^{b,c}$	0.4453	0.4447
pro	0.4281^{b}	0.3865	0.4272
avg	0.3636	0.3276	0.3309
w-avg	0.3575	0.3218	0.3214
m-ratio	0.6873	0.6661	0.6674

Table 3: Diagnostic evaluation results on union alignments

5.2 Automatic Metrics

We also evaluated the performances of the MT systems using a set of state-of-the-art automatic evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). Table 5 presents the system-level evaluation results for the different types of metrics considered (automatic, diagnostic and human judgements). For diagnostic evaluation it reports the weighted averages (see w-avg in Tables 1, 2, 3 and 4). According to BLEU, NIST and METEOR, Google

	Systems		
	Google	Moses	Systran
a	$0.5126^{b,c}$	0.4365	0.4365
n	$0.5494^{b,c}$	0.5042	0.4989
v	0.4074^{c}	0.4261^{c}	0.3496
r	0.5768	0.5431	0.5603
dt	$0.6529^{b,c}$	0.5926	0.6248
misc	$0.7195^{b,c}$	0.6628	0.6542
pro	$0.6331^{b,c}$	0.5493	0.6030
avg	0.5788	0.5307	0.5325
w-avg	0.5683	0.5285	0.5183
m-ratio	1.0926	1.0940	1.0764

Table 4: Diagnostic evaluation results on intersection alignments

is the best system, followed by Moses and Systran, while TER ranks Systran over Moses. Diagnostic evaluation on gold standard alignment also yields the same ranking as BLEU, NIST and METEOR. More importantly, for this work, the use of any automatically derived word alignments (i.e., grow-diag-final, union or intersection) in diagnostic evaluation replicates the same ranking obtained with gold standard alignments.

Method		Systems		
		Google	Moses	Systran
Diagnostic	manual	0.5201	0.4831	0.4815
	gdf	0.3816	0.3445	0.3431
	union	0.3575	0.3218	0.3214
	intersection	0.5683	0.5285	0.5183
Automatic	BLEU	0.2012	0.1621	0.1471
	NIST	5.11	4.54	4.44
	METEOR	0.5033	0.4390	0.4258
	TER	0.6508	0.7059	0.6980
Human	Evaluator 1	3.7864	3.3658	3.4497
	Evaluator 2	4.2417	3.9989	4.0503

Table 5: System level evaluation results

5.3 Human Judgements

Tables 6 and 7 present the results of human evaluation of the MT systems. The mean of adequacy (adq) and fluency (fln) is considered as the overall human judgement score. According to both evaluators, at the system level, Google is the best system, followed by Systran and Moses. As far as fine-grained checkpoint-specific

human judgements are concerned, both evaluators agree that Google handles all of the linguistic phenomena better than the other two systems, as is also revealed by the diagnostic evaluation. According to evaluator 1, Moses translates nouns better than Systran does, while Systran does well on adjectives, verbs, adverbs and determiners. Diagnostic evaluation matches almost perfectly with fine-grained checkpoint-specific human judgements obtained from evaluator 1, except for the translation of verbs for Moses and Systran. But, according to evaluator 2, Moses only translates determiners better than Systran, while Systran does better on the rest of the checkpoints.

	Google	Moses	Systran
Adq	3.9284	3.4765	3.5906
Fln	3.6443	3.2550	3.3087
Avg (Adq, Fln)	3.7864	3.3658	3.4497
noun	4.4758	4.0435	4.0097
verb	4.2138	3.9430	4.1900
adverb	4.6171	4.3237	4.4138
adjective	4.2296	3.8163	4.0302
determiner	4.7754	4.4727	4.7578

Table 6: Human judgements of evaluator 1

	Google	Moses	Systran
Adq	4.6578	4.6577	4.6711
Fln	3.8255	3.3400	3.4295
Avg (Adq, Fln)	4.2417	3.9989	4.0503
noun	4.4734	4.2302	4.3318
verb	4.4143	4.3868	4.4043
adverb	4.6507	4.4855	4.4908
adjective	4.6324	4.4542	4.5210
determiner	4.3605	4.0937	4.0047

Table 7: Human judgements of evaluator 2

Table 8 presents Pearson's correlations for checkpoint-specific evaluation across systems. It shows that checkpoint-specific diagnostic evaluation using either manual or automatic alignments correlates well with checkpoint-specific human judgements in general. However, the correlation is very poor in the case of verbs, as both human evaluators preferred Systran over Moses, while diagnostic evaluation (even with gold standard alignments) ranked Moses over Systran for this checkpoint. We manually inspected the outputs of Moses and Systran for those sentences for which diagnostic evaluation contradicted human evaluation for the verb checkpoint. We found that in some of the cases the problem was due to the failure of DELiC4MT to consider synonyms. Most of the existing automatic evaluation metrics (except METEOR) also suffer from this problem. Availability of mutiple reference translations can circumvent this problem and DELiC4MT also supports evaluation with mutiple references. It should be also noted that the scoring of DELiC4MT being n-gram based, the metric might be slightly biased toward SMT systems (Callison-Burch et al, 2006).

The checkpoint-specific inter-annotator agreements (Fleiss' Kappa) between the two annotators were 0.32 (adjectives), 0.13 (adverbs), 0.12 (determiners), 0.24 (nouns) and 0.29 (verbs). This somewhat low agreement may be due to the fact that although the evaluators are experienced in translation evaluation in terms of adequacy and fluency, they never performed diagnostic evaluation of this sort. It can be noticed from Tables 6 and 7 that evaluator 2 consistently gives higher scores for adequacy and fluency than evaluator 1 across systems; but these scores still correlate perfectly (cf. Table 9). A limitation of the current study regards the low number of human annotators, as having more of them might probably result in more stable results.

Finally, Table 9 presents the Pearson's correlation coefficients between the system-level scores across systems. As it can be seen from this table, system-level diagnostic evaluation scores obtained on automatically derived word alignments correlate very highly with those obtained on the gold standard alignment. In fact, diagnostic evaluation using grow-diag-final and union alignments (as opposed to using manual alignments) was found to correlate better with human judgements, while the use of intersection alignments produced better correlations with the majority of the automatic MT evaluation metrics. This indicates that using automatic word alignments is sufficient for carrying out diagnostic evaluation. Diagnostic evaluation correlates well with all automatic evaluation scores (including TER, which being an error metric shows strong negative or inverse association) as well as human judgements, indicating that this type of evaluation is accurate at predicting true system quality.

Pearson's Correlation	Noun	Verb	Adv	Adj	Det
Evaluator 1 – Evaluator 2	0.880	0.959	0.962	0.987	0.327
Evaluator 1 – Diagnostic (manual)	0.999	-0.305	0.951	0.872	0.885
Evaluator 2 – Diagnostic (manual)	0.898	-0.021	0.830	0.940	0.729
Evaluator 1 – Diagnostic (gdf)	0.999	-0.123	0.890	0.710	0.873
Evaluator 2 – Diagnostic (gdf)	0.853	0.163	0.980	0.815	0.746
Diagnostic (manual) – Diagnostic (gdf)	0.996	0.983	0.704	0.964	1.000
Diagnostic (manual) – Diagnostic (union)	0.999	0.968	0.704	0.984	1.000
Diagnostic (manual) – Diagnostic (intersection)	0.998	0.932	0.996	0.999	0.999

Table 8: Pearson's correlation for checkpoint-specific evaluation across systems

		Diagnostic				Human	
		manual	gdf	union	intersection	evaluator 1	evaluator 2
Diagnostic	manual	1.000	1.000	1.000	0.988	0.975	0.972
	gdf	1.000	1.000	1.000	0.987	0.976	0.973
	union	1.000	1.000	1.000	0.983	0.980	0.978
	intersection	0.988	0.987	0.983	1.000	0.927	0.922
Automatic	BLEU	0.972	0.971	0.966	0.997	0.895	0.890
	NIST	0.995	0.994	0.992	0.998	0.947	0.942
	METEOR	0.992	0.992	0.989	0.999	0.940	0.935
	TER	-0.986	-0.986	-0.990	-0.947	-0.998	-0.998
Human	Evaluator 1	0.975	0.976	0.980	0.927	1.000	1.000
	Evaluator 2	0.972	0.973	0.978	0.922	1.000	1.000

Table 9: Pearson's correlation between the system level scores

6 Conclusions and Future Work

This paper has evaluated a diagnostic metric that assesses the performance of MT systems on user-defined linguistic phenomena. This has been done by means of a case study for the English–French language direction.

As this metric is dependent on word alignments, one of the objectives was to find the margin of error in diagnostic evaluation using automatic word alignments as opposed to using gold standard manual alignments. In order to determine that, we carried out diagnostic evaluation using manual alignments as well as a set of commonly used automatic alignments (grow-diag-final, union and intersection). In addition, we also calculated the correlation with several state-of-the-art automatic MT evaluation metrics as well as with human judgements.

From the experimental results we found that automatically-derived word alignments can be considered as effective as gold standard alignments when carrying out diagnostic evaluation. We also observed that diagnostic evaluation can accurately capture translation quality and, in general, correlates well both with system-level automatic evaluation metrics and with human judgements.

As an extension to this work, we would like to explore the impact of different automatic aligners on the results of diagnostic evaluation of MT. Also, the low correlation with human judgements obtained for verbs requires a deeper analysis of this linguistic phenomenon and how it is treated by the diagnostic metric, which we plan to explore in further detail.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreements FP7-ICT-4-248531 (CoSyne) and PIAP-GA-2012-324414 (Abu-MaTran) and through Science Foundation Ireland as part of the CNGL (grant 07/CE/I1142).

References

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluation the Role of Bleu in Machine Translation Research. In *Proceedings of EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145.

Fishel, M., Bojar, O., and Popović, M. (2012). Terra: a collection of translation error-annotated corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* , pages 7–14.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86.

LDC (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report. Revision 1.5.

Max, A., Crego, J. M., and Yvon, F. (2010). Contrastive lexical evaluation of machine translation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The blinker project. *CoRR*, cmp-lg/9805005.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2* , pages 1086–1090.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* , pages 311–318.

Popović, M. (2011). Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Popović, M. and Burchardt, A. (2011). From human to automatic error classification for machine translation output. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*.

Popović, M., Ney, H., de Gispert, A., Mariño, J. B., Gupta, D., Federico, M., Lambert, P., and Banchs, R. (2006). Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 1–6.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, Department of Computer and Information Science, University of Pennsylvania. (3rd revision, 2nd printing).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Stolcke, A. (2002). SRILM–an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

Toral, A., Naskar, S. K., Gaspari, F., and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. In *The Prague Bulletin of Mathematical Linguistics*, 98:121–132.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., and Zhao, T. (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1.* , pages 1121–1128.

Towards a Generic Approach for Bilingual Lexicon Extraction from Comparable Corpora

Dhouha Bouamor
CEA, LIST, Vision and Content Engineering Laboratory, 91191 Gif-sur-Yvette CEDEX France
dhouha.bouamor@cea.fr

Nasredine Semmar
CEA, LIST, Vision and Content Engineering Laboratory, 91191 Gif-sur-Yvette CEDEX France
nasredine.semmar@cea.fr

Pierre Zweigenbaum
LIMSI-CNRS, F-91403 Orsay CEDEX France
pz@limsi.fr

Abstract

This paper presents an approach that extends the standard approach used for bilingual lexicon extraction from comparable corpora. We focus on the problem associated to *polysemous words* found in the seed bilingual lexicon when translating source context vectors. To improve the adequacy of context vectors, the use of a WordNet-based Word Sense Disambiguation process is tested. Experimental results on four specialized French-English comparable corpora show that our method outperforms two state-of-the-art approaches.

1 Introduction

Bilingual lexicons play an important role in many natural language processing applications such as machine translation or cross-language information retrieval (Shi, 2009). Research on lexical extraction from multilingual corpora have largely focused on parallel corpora. The scarcity of such corpora in particular for specialized domains and for language pairs not involving English pushed researchers to investigate the use of comparable corpora (Fung, 1998; Rapp, 1995; Chiao and Zweigenbaum, 2003), in which texts are not exact translation of each other but share common features.

The basic assumption behind most studies is a *distributional* hypothesis (Harris, 1954), which states that words with a similar meaning are likely to appear in similar contexts across languages. The so-called ***standard approach*** to bilingual lexicon extraction from comparable corpora is based on the characterization and comparison of lexical environments represented by *context vectors* of source and target words. In order to enable the comparision of source and target vectors, words in the source vectors are translated into the target language using an existing bilingual dictionary.

The core of the standard approach is the bilingual dictionary. Its use is problematic when a word has several translations, whether they are synonymous or polysemous. For instance, the French word *action* can be translated into English as *share, stock, lawsuit* or *deed*. Identifying which translations provided by a given bilingual dictionary are most relevant impacts the quality of the extracted bilingual lexicons. The standard approach considers all available translations and gives them the same importance in the resulting translated context vectors independently of the domain of interest and word ambiguity. For instance, in the financial domain, translating *action* into *deed* or *lawsuit* would introduce noise in context vectors.

In this paper, we present a novel approach which addresses the word polysemy problem neglected in the standard approach. We exploit a Word Sense Disambiguation (WSD) process that identifies the translations of polysemous words that are more likely to give the best representation of context vectors in the target language. For this purpose, we employ five WordNet-based semantic relatedness measures and use a *data fusion* method that merges the results obtained by each measure. We test our approach on four specialized French-English comparable corpora \\\\\\\(*financial, medical, wind energy and mobile technology)* and report improved results compared to two state-of-the-art approaches.

The remainder of this paper is organized as follows: the next section describes the standard ap-

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 143–150.