Arabic Dialect Handling in Hybrid Machine Translation

Hassan Sawaf
AppTek Inc.
6867 Elm Street #300
McLean, VA 22101
hassan@apptek.com

Abstract

In this paper, we describe an extension to a hybrid machine translation system for handling dialect Arabic, using a decoding algorithm to normalize non-standard, spontaneous and dialectal Arabic into Modern Standard Arabic. We prove the feasibility of the approach by measuring and comparing machine translation results in terms of BLEU with and without the proposed approach. We show in our tests that on real-live broadcast input with transcriptions of dialectal speech we achieve an increase on BLEU of about 1%, and on web content with dialect text of about 2%.

1 Introduction

In comparison with broadcast news speech uttered by anchor speakers, spontaneous speech poses additional difficulties for the task of machine translation (MT). Typically, these difficulties are caused by the lack of conventional syntactic structures because the structures of spontaneous speech differ from that of standard written language.

This paper is organized as follows: We describe briefly the utilized approach of hybridization to machine translation for general text input in Section 2. Then we describe the approach on dialect handling and normalization for dialectal and noisy text in Section 3. In Section 4 we present some experiments and analyze the results. We conclude this paper in Section 5.

2 Hybrid Machine Translation

Our prime motivation for utilizing a hybrid machine translation system is to take advantage of the strengths of both rule-based and statistical approaches, while mitigating their weaknesses.

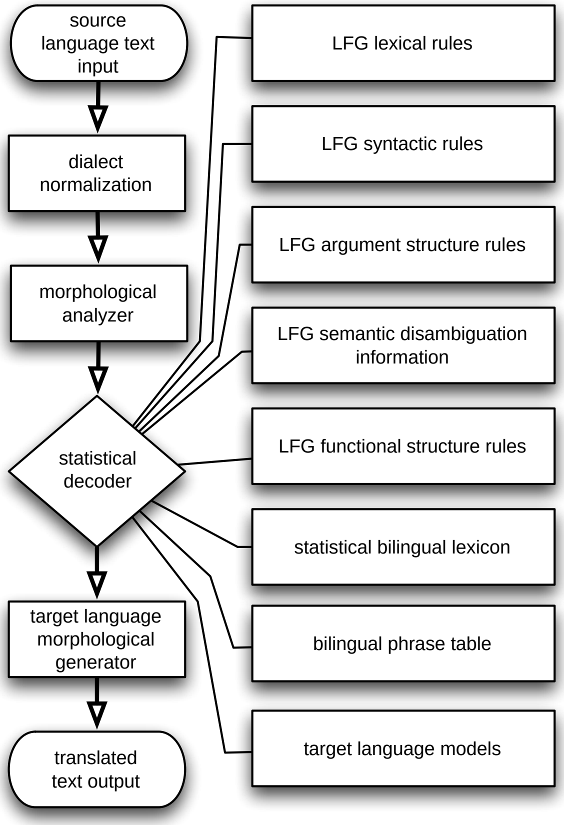
source language text input
dialect normalization
morphological analyzer
statistical decoder
target language morphological generator
translated text output
LFG lexical rules
LFG syntactic rules
LFG argument structure rules
LFG semantic disambiguation information
LFG functional structure rules
statistical bilingual lexicon
bilingual phrase table
target language models

Figure 1. Flow diagram of main components of the dialect normalization process.

Thus, for example, we want a rule that covers a rare word combination or phrasal construction to take precedence over statistics that were derived from sparse data (and thus not very reliable).

For that reason, we see hybridization of machine translation as an advanced form of smoothing for a statistical machine translation.

Additionally, rules covering long-distance dependencies and embedded structures should be weighted favorably, if relevant to a translation approach, since these constructions are more difficult to process in statistical MT.

Conversely, we would like a statistical approach to take precedence in situations where word combinations and phrasal structures occur in training in an amount to estimate reliable statistics.

An aspect that is extremely important for further processing the MT output, is the weakness which statistical MT sometimes has in "informativeness" (accuracy in translation, with special regards to information content) due to the high influence of the target-language model. Single words that may make a disproportionately heavy contribution to informativeness, such as terms indicating negation or important content words may be missing, for example, or adjectives that are misplaced. This phenomenon occurs mostly in cases, where this content word occurs in a rare context, whereas the context without that content word is seen more frequently.

In our approach to HMT, the statistical search process has access to the complete information database available in the rule-based engine, as outlined in Figure 1. The components in the figure will be briefly described in the following subsections of this paper.

Statistical Machine Translation is traditionally represented in the literature as choosing the target (English) sentence $\boldsymbol{e} = e_1...e_I$ with the highest probability given a source (French) sentence $\boldsymbol{f} = f_1...f_J$:

$$\hat{\boldsymbol{e}} = argmax_{\boldsymbol{e}}\{Pr(\boldsymbol{e}|\boldsymbol{f})\} \qquad (1)$$

The rich syntactic/semantic information is derived from the rule-based engine parser that produces syntactic trees annotated with rich semantic and syntactic annotations.

The hybridization is then accomplished by treating all the pieces of information as feature functions in a log-linear framework:

$$Pr(\boldsymbol{e}|\boldsymbol{f}) = p_{l1..M}(\boldsymbol{e}|\boldsymbol{f}) =$$

$$\frac{exp[\sum_{m=1..M} l_m\, h_m(\boldsymbol{e},\boldsymbol{f})]}{\sum_{\boldsymbol{e}'} exp[\sum_{m=1..M} l_m\, h_m(\boldsymbol{e}',\boldsymbol{f})]}\ ; \qquad (2)$$

we obtain the following decision rule out of (1):

$$\hat{\boldsymbol{e}} = argmax_{\boldsymbol{e}}\{Pr(\boldsymbol{e}|\boldsymbol{f})\} =$$

$$argmax_e\{\sum_{m=1..M} l_m\, h_m(\boldsymbol{e},\boldsymbol{f})\}\ . \qquad (3)$$

Incorporation of these different knowledge sources (rule-based and statistical) is then achieved by adding feature functions to the criterion, and allowing a training algorithm like Generalized Iterative Scaling (GIS) or Improved [generalized] Iterative Scaling (IIS) to train the weights of the feature in context to the other features in respect to the final translation quality measured by an error criterion (Och and Ney, 2002).

2.1 Arabic Preprocessing and Segmentation

Adequate preprocessing and segmentation is very helpful in processing Arabic text. This paper will elaborate in detail on the preprocessing of the textual input in the next section. The goal of the preprocessing is to remove any textual noise out of the input text, and to process the input text in such a way, that it matches the data which the main MT system was trained on.

Following to the preprocessing, the text is morphologically processed and segmented.

For languages like Arabic, morphology plays a big role. For tasks to recognize, translate and process Arabic Broadcast News with an original vocabulary size of more than 600K words, the morphology module can decrease the 600,000 word vocabulary by up to 70 percent, reducing it to 256,000 morpheme-units without losing any relevant information. The biggest advantage of reducing the vocabulary size is that the amounts of so-called “unknown words” (i.e. words never observed in historical data) can be reduced, as the system can find more morphological derivatives for a new (i.e. historically unfound) word and can therefore relate this new word to a translation by using morphologically related words.

2.2 Rule-Based Models

For higher “informativeness” our rule-based module feeds many different information streams into the core decoder. For our rule-based module, we employ a Lexical Functional Grammar (LFG) system ((Kaplan and Bresnan, 1982), (Shihadah and Roochnik, 1998)).

The LFG system incorporates a richly-annotated lexicon containing functional and semantic information. It also produces richly-annotated intermediate outputs (e.g. phrasal parses) that is processed by the decoding algorithm:

- **Source language c-structure** (or "constituent structure") – a phrase-structure tree;
- **Part of Speech** as in noun, verb, adjective, etc.;
- **Word/phrase order**;
- **Source language f-structure** (or "functional structure") – a Directed Acyclic Graph (DAG) containing attribute-value pairs that specify, for example, grammatical information such as subject/object and case (genitive objective, accusative, dative and others depending on language), argument structure (predicate, argument, adjunct), semantic disambiguation information (human, animate, concrete, etc.) and grammatical information including gender, plurality, mood, tense, aspect, polarity, speech act and other information;
- **Target language f-structure** – an f-structure that has been modified and restructured to enable generation of target-language text;
- **Lexical entries** contain lexically-determined grammatical information of the type listed in the f-structure above.

2.3 Functional Models

We use functional constraints for lexical information in source and target text, which performs a deeper syntactic and semantic analysis on the translation to result in more accurate translations with greater contextual relevance. Functional constraints are multiple, and some of these functions are language dependent (e.g. gender, polarity, mood, etc.).

While most functional relations are word-based, some functions can be across languages or within a specific language. For instance, words such as "man," "woman," and "president," are generally "human," therefore the function "human" is considered "positive." However, there are other words such as "manager," "caller," and "driver" that can be a "human positive" but they also could be "human negative," for example if the "driver" is a printer driver and not a human. These concepts depend on the semantic and syntactic environment.

An example for a language-specific function could be gender, as objects can have different genders in different languages (e.g. "table" in English is neuter, "Tisch" in German is masculine, "table" in French is feminine, and "طـاولـة" in Arabic is feminine: all four words are translations of each other).

The parse trees that feed the statistical module, make use of all these semantic attributes, as well as the syntactic features.

2.4 Statistical Translation Models

The Statistical Machine Translation approach we chose for conducting the experiments is a phrase-based approach similar to the alignment template approach described in (Och and Ney, 2004). Compared to traditional word-based statistical MT systems, these methods have the advantage of a capability to learn translations of phrases, not just individual words, which permits it to encompass the functionality of example-based approaches and translation memories. Using the Maximum Entropy approach as described in (Och and Ney, 2004) the other advantage is that it allows for the combination of many knowledge sources, by framing them as feature functions that are combined using a Maximum Entropy framework.

The translation models introduced for the system that is described here is a combination of statistically learned lexicons and statistically learned phrase tables (Koehn et al., 2007), (Och and Ney, 2004).

2.5 Statistical Language Models

Our MT uses a combination of standard n-gram language models and structural language models analogous to the work of ((Sawaf et al., 2000), (Charniak et al., 2003)). In order to improve MT quality, language model feature functions are introduced, where the language model feature functions are, for example, combinations of standard word-based 5-gram models, POS-based 5-gram models, and a time-synchronous CYK-type parser.

3 Dialect Normalization

To translate documents from broadcasts and the Internet, (e.g. blogs and news sites, emails, speech transcripts, etc.) the need for noise reduction by normalization is critically important to accuracy. Most of these transcripts are not only in MSA but include also non-standardized transcripts of dialect words, which can create inconsistencies between the documents that need to be translated and the model components inside the statistically trained machine translation system. To cope with this problem, machine learning techniques are utilized to “normalize” those into a format that is consistent with the material from which the statistical and corpus-based components are built. The process is basically a specialized monotone translation with local reordering, allowing input phrases to be translated, so that special email, chat, or spoken jargon can be translated to standard text, and misspellings corrected. For this, we can use a standard statistical approach, and the use of a strong background lexicon to achieve good results in the spell checking.

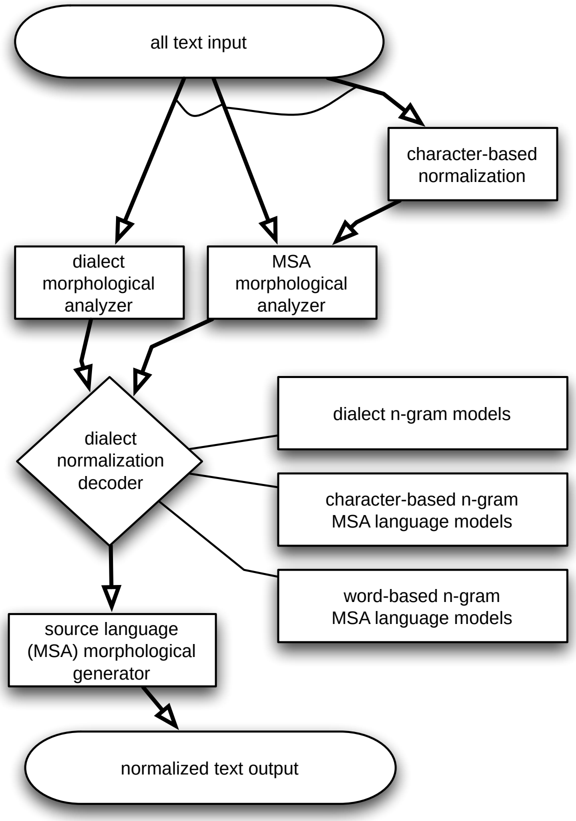
all text input
character-based normalization
dialect morphological analyzer
MSA morphological analyzer
dialect normalization decoder
dialect n-gram models
character-based n-gram MSA language models
word-based n-gram MSA language models
source language (MSA) morphological generator
normalized text output

Figure 2. Flow diagram of main components of the dialect normalization process.

Transferring the words from the local dialects to MSA can be seen as an Interlingua approach for dialect normalization. Due to the nature of dialects, usually some information can be missing, that should be available for MSA, e.g. case endings, time information, and sometimes gender. These information can usually be inferred by the context which is modeled by a strong language model for MSA, but generally might still form a weak spot in the word sequence for the following MT system.

3.1 Proposed Approach for Dialect Normalization for Arabic

The above-mentioned process and technique can be used for dialect normalization and translating Arabic dialects into MSA.

We can describe mathematically the dialect normalization procedure as the following mathematical representation:

$$\boldsymbol{f} = argmax_{\boldsymbol{f}}\{Pr(\boldsymbol{f}|\boldsymbol{F})\} \; ; \qquad (4)$$

where $\boldsymbol{f}$ denotes the optimal MSA, given a sequence of Arabic words $\boldsymbol{F}$ which contain dialect words from various dialects. The decomposition of the general problem into smaller models is analogous to the machine translation task and we obtain two main models, which are tightly interwoven:

- **Dialect Analysis and Processing** $Pr(\boldsymbol{f},\boldsymbol{F})$ - this process generates all possible MSA word and phrase sequences that can be related to the original sentence; and
- **MSA Generation** $Pr(\boldsymbol{f})$ - this process evaluates the sequence of generated MSA words and weighs them according to the context within the generated sentence.

Figure 2 shows an outline of the process. The different components used in the process are described in the following sub-sections in detail.

3.2 Dialect Analysis and Processing

The preprocessing of the input text, that might contain dialect text is processed in three parallel ways before the main decoding step:

- by a *character based dialect normalization*, which utilizes simple rules to convert words into the most similar MSA word, as seen in the training data of the MT system[1]. These simple rules can be hand-coded and enriched by rules which can be learned from bi-dialectal[2] with bilingual

[1] Please refer to Section 4 for details on the training data, vocabulary size of background lexicon.

[2] A sentence-aligned corpus with each sentence in two different dialects, e.g. MSA and Lebanese Arabic.

alignment modeling software like GIZA++ (Al-Onaizan et al. 1999), and phrasal extraction modules, as described in (Och and Ney, 2004) and (Koehn et al., 2007). There is a separate training process to generate dialect Arabic/MSA word and phrase translation pairs for each dialect. This training is carried out on the bi-dialectal corpus, leaving out a small subset of sentences for parameter optimization and test. For further processing, all possible permutations and possibilities of transliteration of each word are taken into consideration, weighted by the grade of transformation from the original form. This step allows processing of Arabic which is encoded in latin characters, e.g. "tawle" for "طاولة", i.e. English "table".

- by a *non-dialect morphological analyzer*, which is a standard, FSA-based MSA morphological analyzer as described in (Köprü and Miller, 2009). All different morphological derivations are taken into consideration for further processing.
- by a *dialect-specific morphological analyzer*. This process utilizes a morphological set of hand-crafted rules which describe the general morphology of the different dialects described above. In addition to the usual word segmentation, each word is tagged with the dialect that the firing rules are written for. This yields into potential multiple output with the same morphological analysis, but with different tags (e.g. Jordanian and Syrian for the word "جـــاي", i.e. English "is approaching" or "I am coming!").

As part of the analysis process, a class-based n-gram language model, where the classes specify the dialects is utilized. This class-based n-gram language model is analogous to a part-of-speech based language model and can be denoted as:

$$Pr(\boldsymbol{f}) =$$

$$argmax_e\{\prod_{j=1..J} Pr(f_j \mid c_j)\, Pr(c_j \mid c_{1...j-1}\} \; ; \quad (5)$$

Where f_j denotes the MSA word in position j, c_j denotes the dialectal class. Currently, we distinguish 16 different main Arabic dialects. These dialects consist of MSA and 15 colloquial Arabic dialects from the following regions:

- East Arabic: Lebanese, North Syria, Damascus, Palestine, Jordan;
- Gulf Arabic: Northern Iraq, Baghdad, Southern Iraq, Gulf, Saudi-Arabia, Southern Arabic Peninsula;
- Nile Region: Egypt and Sudan;
- Maghreb Arabic: Libya, Morocco, Tunisia.

At the current state, the statistics for the word-dialect probability $Pr(f_j \mid c_j)$ is estimated by the counts $N(f_j\,,\,c_j)$ and $N(c_j)$ from the dialect-tagged bi-dialectal corpora available. The dialect-sequence probability $Pr(c_j \mid c_{1...j-1}\}$ is chosen to be as follows:

$$Pr(c_j \mid c_{1...j-1}) =$$

$$= \frac{1}{j} \sum_{k=1..j} \partial(c_j, c_k) \;\; : \qquad iff\; c_j = c_{j-1} \quad , \qquad (6)$$

$$= \frac{1}{j \cdot \#(\zeta(c_j))} \sum_{k=1..j} \partial(c_j, c_k) \;\; : \;\; iff\; c_j \in \zeta(c_j) \;\; ,$$

$$= \frac{1}{J \cdot \sum_c \#(\zeta(c))} \quad : \qquad else \;\; ;$$

where $\partial(\cdot\,,\cdot)$ is the Kronecker delta, which is 1 if the arguments are equal, $\zeta(c_j)$ results the set which includes the dialect class c_j (e.g. the class c_j = 'Lebanese' would result in $\zeta(c_j)$ = 'East Arabic'). $\#(\zeta(c_j))$ denotes the size of the dialect set $\zeta(c_j)$.

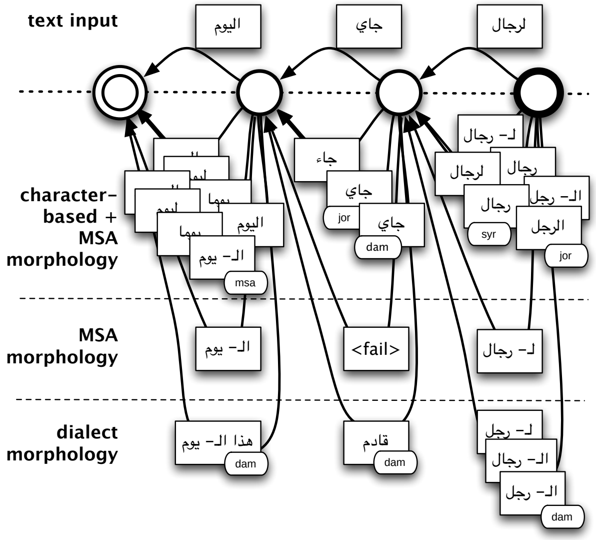

Figure 3. Example sentence and possible result graph for pre-processing. The sentence would translate in Syrian (syr) or Damascene (dam) Arabic "the man is approaching today". The second input word cannot be analyzed by the MSA morphological analyzer. the MSA morphological analyzer fails to analyze the second token, as it is a distinct dialect word "approach, come".

This way of modeling results in a preference, that words of the same dialect are most likely to follow each other, words of similar dialects (i.e. in the same dialect family) are less likely to follow each other, and the least probable is that dialects of different dialect families follow each other.

All three processing steps generate a graph of possible word sequences. Figure 3 shows an example sentence and the resulting graph. For easier visualization, the probabilities $Pr(f_j,\boldsymbol{F})$ for each hypothesis is not shown in the figure, but for the process these values are kept. All annotation (e.g. POS, stem, gender, case, etc.) which are found on each of the processing steps are kept, to potentially be used by the following MT system, as far as the MT system can process the information.

3.3 MSA Generation

The decision process on the selection of dialect normalized words takes into account a character-based n-gram language model, a word-based n-gram language model for the normalized MSA output. For the word-based language model, we chose a word 5-gram with Kneser-Ney smoothing (Kneser and Ney, 1995), and for the character-based language model, we chose a 20-gram backing-off language model.

3.4 Dialect Normalization Decoding

The decoding algorithm for dialect normalization is a beam search decoding algorithm. As the alignment between dialect Arabic and MSA is largely monotone, we constrained the reordering in the decoder to be very local, i.e. within a range of 3 words. Experiments show, that this range is a good balance between speed, memory utilization and end-to-end translation quality.

As many of the Arabic input words can be either MSA or a dialect word at the same time, the use of networks as input into the decoder from the preprocessing step instead of single-best word sequences show to produce the best results.

For the presented translation task, a bi-dialectal corpus is used which consists of sentences in different Arabic dialects and Modern Standard Arabic.

3.5 Dialect Normalization for HMT Training

We performed the dialect normalization in two modes: we ran the dialect normalization just on new test data, and we ran the dialect normalization on training and test data. In Section 4 we show our experimental results.

4 Experiments

We used the evaluation score BLEU (the higher the better) to evaluate our approach (Papineni et al., 2002). For optimal results, the machine translation systems were using adaptation techniques, where portions of the training corpus get weighted higher, where the language model perplexity on the test sentences are very low. The multiplier used was a factor of 10x, and the sub-corpus with the 700K words where the perplexity is minimal are used as an adaptation corpus.

	MT Train		**MT Test**	
	BC/BN	**Web**	**BC/BN**	**Web**
# sentences	14,.3M	38.5K	12,4K	547
# words	375M	816.3K	132.6K	18.6K
Vocabulary	256K		N/A	
OOV	N/A		0.2%	0.3%

Table 1. Corpus Statistics for Training and Test of both the SMT and the HMT systems, respectively. Web test corpus is NIST MT08 WB portion.

Table 1 shows the corpus statistics for the SMT and HMT training. The broadcast test data were collected by a service provider in the United Arab Emirates, and translated to evaluate MT systems for their use. The data was collected in such a way, that almost all dialects of the whole Arabic speaking region is covered in the test corpus by integrating not only newscasts, but also interviews, films and TV shows. The training data are the data which can be obtained from LDC, as well as additional training data that were collected in a semi-automatic way by using the machine translation and human post-editing.

Table 2 shows the corpus statistics for Training and optimization of the dialect normalizer, using bi-dialectal data.

	DNorm Train		DNorm Dev	
	BC/BN	Web	BC/BN	Web
# sentences	271K	1.6M	5K	8K
# words	3.1M	28M	57.3K	111K

Table 2. Corpus Statistics for Training and development of the dialect normalization (DNorm).

To compare the use of dialect normalization on the different MT approaches, we are showing results with and without dialect normalization in Table 3. LFG denotes a rule-based MT system, SMT is a state-of-the-art phrase-based MT, and HMT is the described hybrid MT system, which incorporates both the described SMT and the LFG MT subsystems.

BLEU	LFG	SMT	HMT
No DialectNorm	18.1%	35.4%	37.3%
With DialectNorm	19.5%	36.4%	38.5%

Table 3. Effect of Dialect Normalization on three different MT approaches.

Table 4 shows the results of dialect normalization ran only on the MT test data compared to MT without dialect handling, and compared to tests, where both the training corpus and the test corpus has been processed by the dialect normalization.

BLEU on HMT	BC/BN	Web
No DialectNorm	35.5%	39.9%
DialectNorm in Test	35.9%	40.5%
DialectNorm in Test & Train	36.4%	42.1%

Table 4. Effect of Dialect Normalization ran on test corpus only and consistently on test and training corpus.

A sentence of the test corpus ran through the different MT approaches can be seen in Table 5. ICA denotes the original sentence, and MSA is the dialect normalized version of the same sentence.

The increase of quality measured by BLEU is clear both on the broadcast tests as well as on the web text tests. For the broadcast experiments, the dialect normalization increases the quality by an absolute BLEU of 0.4% compared to the non-normalized baseline by normalizing the test set, and by normalizing both training and test corpus, the system increases another absolute 0.5%, to reach a BLEU of 36.4%. For Web content, the achievement is even higher: running the normalization only on the test portion increases the BLEU by an absolute 0.5%, running it on training and test increases the BLEU by another absolute 1.6% on the MT08wb test set, leading to 42.1%.

ICA	خلَالْ الْثْنَعشْ سَنَة الِّي فَاتَتْ انْكتَلْ أَزْيَدْ من أَلْفُ وُمَيةْ صَحَفَي وُعَاملْ اَبْ مَجَالْ الْاَعْلَامْ مِنْ جَانَوْا يِأَدُّونْ وَاجِبْهُمْ، انْكَتِلَوْا لَأَنْ فَدْ وَاحِدْ يِمْكَنْ مَا عَجَبَهْ الِّي يِكْتِبُوهْ أَوْ الِّي يِكُولُوهْ أَوْ لَأَنْ هُمَّهْ جَانَوْا اِبْ مُكَانْ وُ وَكِتْ مَا جَانْ لَازِمْ يِكُونُونْ مَوْجُودِينْ بِيهْ.
MSA	خلَالْ الإثني عشر سَنَة الذي فَاتَتْ انقتل أَزْيَدْ مِنْ أَلْفُ ومئة صَحَفِي وُعَامِلْ في مَجَالْ الْاِعْلَامْ مِنْ كانوا يِأَدُّونْ وَاجِبْهُمْ، انقتلوا لَأِنْ هناك وَاحِدْ يِمْكِنْ مَا عَجَبَهْ الذي يِكْتِبُوهْ أَوْ الذي يقولوه أَوْ لَأِنْ هم كَانوا في مُكَانْ وُ وقت مَا كان لَازِمْ يِكُونُونْ مَوْجُودِينْ فِيه.
LFG	During ethnic year 10 which passed killed from one thousand and one hundred journalists and a factor increase. In the media domain who were performing duty killed, there one can. What astonishment which write or which say or because they were in place. And timed were a necessary existing in.
SMT	During the 12 years which passed died more than one thousand and one hundred journalists and worker in the media domain who were performing duty, died because there one can not like which write or say or they were in place and a time not were a necessary existing in.
HMT	During the twelve years which passed more than one thousand and one hundred journalists and media workers performing their duty, were killed because there was one who did not like what they wrote or what they said or because they were in a place and time they should not be present at.

Table 5. Example sentence Arabic Broadcast News; Original in Iraqi Colloquial Arabic (ICA) and Modern Standard Arabic (after normalization; MSA) and the translations using rule-based MT (LFG), statistical MT (SMT) and hybrid MT (HMT).

5 Conclusion

In this paper, we introduced a MT system, that is optimized to handle dialect, spontaneous and noisy text from broadcast transmissions and web content. We compared three approaches of MT with this task, and showed that a hybrid approach performs best out of these.

We also described a novel approach on how to deal with Arabic noisy and dialectal data by normalizing the input text to a common form, and then processing this. By processing the training and the test corpora, we could improve the translation quality by about absolute 2% for Web text and about 1% absolute for broadcast transmissions.

For the future, we would like to investigate the automatic learning of the dialect similarity/ transition probability and test this in regards to the MT quality. We also would like to see if we can combine the dialect normalization into the actual MT process. Furthermore it needs to be investigated, whether the manual dialect classification is superior to automatic clustering algorithms, based on grapheme features and/or phoneme features.

References

Y. Al-Onaizan, Jan Curin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Purdy, N. A. Smith, and D. Yarowsky. 1999. "Statistical Machine Translation: Final Report," Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roosin. 1990. A Statistical Approach to Machine Translation. Computational Linguistics, 16, pp. 79–85. Cambridge, MA.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2), pp. 263–311. Cambridge, MA.

E. Charniak, K. Knight, and K. Yamada. 2003. Syntax-Based Language Models for Statistical Machine Translation. In Proceedings of MT Summit IX, pp. 23–27. New Orleans, LA.

R. Kaplan and J. Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In The Mental Representation of Grammatical Relations, pp. 173–281. MIT Press.

R. Kneser, and H. Ney. 1995. Improved backing-off for m-gram language modeling. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1. pp. 181–184.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A Constantin, E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

S. Köpru and J. Miller. 2009. A Unification Based Approach to the Morphological Analysis and Generation of Arabic. CAASL3: Proc. of the 3rd Workshop on Computational Approaches to Arabic-script based Languages. Ottawa, ON, Canada.

F. J. Och, and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In Proceedings of the Ninth Machine Translation Summit, pp. 295–302. New Orleans, LA.

F. J. Och, and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. In Computational Linguistics, 30(4), pp. 417–449. Cambridge, MA.

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

H. Sawaf, K. Schütz, and H. Ney. 2000. On the Use of Grammar-Based Language Models for Statistical Machine Translation. In Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT), pp. 231–241. Trento, Italy.

M. Shihadah, P. Roochnik. 1998. Lexical-Functional Grammar as a Computational-Linguistic Underpinning to Arabic Machine Translation. In Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing. Cambridge, UK.