Using English as a Pivot Language to Enhance Danish-Arabic Statistical Machine Translation

Mossab Al-Hunaity, Bente Maegaard, Dorte Hansen
Center for Language Technology
University of Copenhagen
musab@hum.ku.dk,bmaegaard@hum.ku.dk,dorteh@hum.ku.dk

Abstract

We inspect two pivot strategies for Danish-Arabic statistical machine translation (SMT) system; phrase translation pivot strategy and sentence translation pivot strategy respectively. English is used as a pivot language. We develop two SMT systems, Danish-English and English-Arabic. We use different English-Arabic and English-Danish data resources. Our final results show that SMT systems developed under sentence based pivot strategy outperforms system developed under phrase based pivot strategy, especially when common parallel corpora are not available.

1. Introduction

Developing a statistical machine translation (SMT) system between any two languages usually requires a common parallel corpus. This is used in training the SMT in translating the source language to the target language. Bilingual corpora are usually available for widely spread language pairs like Arabic English Chinese, etc, but when trying to develop SMT systems for languages pair like Arabic-Danish a bilingual corpus unfortunately doesn't exist. The limited data resources make developing SMT for Arabic-Danish a real challenge. To the best of our knowledge, there has not been much direct work on SMT for the Danish-Arabic language pair. Google Translate which is a free web translation service provides the option for translation from Danish to Arabic. Google Translate web service uses gigantic monolingual texts collected by its crawling engine to build massive language models. Aligned bilingual language resources collected through web makes it easy for Google to build SMT between any language pairs. Google performs better between languages pair which has huge common data resources like the case in English and Arabic or English and Chinese. For pairs like Arabic and Danish Google translation quality is quite less than other pairs. A possible explanation for that is the lack of common parallel available resources which control the SMT learning and performance. In our work we don't consider language resources factor alone, but also we concentrate on language specific details like syntax and morphology to tune SMT learning for our Danish-Arabic baseline. We also utilize text processing tools to enhance our baseline performance. Although a parallel corpus is not available for the Danish-Arabic pair, there are lots of parallel English-Arabic and English-Danish resources available. This makes English as a pivot language between Arabic and Danish a favorable choice. Still any language can be used as a pivot language. Our experiments use two separate corpora for Danish–English and English-Arabic SMT systems. Having English as a pivot Language we apply two different pivot strategies:

- Phrase translation pivot strategy.
- Sentence translation pivot strategy.

These methods are based on techniques developed by Utiyama, Isahara (2007), but we apply these techniques with a different perspective. We use non parallel corpora as a source of training data and not corpora with common text. We develop two baselines: Danish-English system that is piped with another English-Arabic system to translate from Danish into Arabic. Each system has different training corpus from the other. Corpora yet share or intercross partially in domain. Languages nature represents another challenge for our baseline. Our System languages are from completely different families which affect experiment results greatly. Another interesting factor is the training data resources. Many previous efforts on SMT systems with pivot language were carried on parallel corpora where data was aligned on sentence level; languages either were from the same nature like European languages Koehn (2009), or they shared a parallel data for the source pivot and target. For example

Habash and Hu (2009) used English as a pivot language between Chinese and Arabic where the three languages in their system were based on the same text. Our work differs in that we train our two systems on two different unrelated sets of data. This is due to the fact of scares parallel data resources between Danish and Arabic. Many pivot strategies are suggested in previous studies like the case with Bertoldi et. al (2008) ,Utiyama, Isahara (2007) and Habash and Hu (2009). We choose to apply our experiments on two strategies; namely phrase translation and sentence translation, due to the available data resources and to hold more control on experiments conditions. We plan to inspect further techniques on Danish Arabic SMT system in future work. Our results show that using English as a pivot language is possible with partially comparable corpora and produces reasonable results. We discover that sentence translation strategy outperforms phrase translation strategy, especially when none parallel or common resources are available. We compare our experiments results with Google Translate to judge system performance. Finally we discuss future research directions we find interesting to enhance our baseline performance. In the next section we describe related work. Section 3 presents our system description. In section 4 we describe our data and present our pivot experiments details. We present our system performance results in section 5. Finally we discuss our conclusions and future work in section 6.

2. Related Work

There has been a lot of work on translation from Danish to English Koehn (2009), and from Arabic to English Sadat and Habash(2006) , Al-Onaizan and Papineni, (2006).Many efforts were spent to overcome the lack of parallel corpora with pivot methods. For example, Resnik and Smith (2003) developed a technique for mining the web to collect parallel corpora for low-density language pairs. Munteanu and Marcu (2005) extract parallel sentences from large Chinese, Arabic, and English non-parallel newspaper corpora. Statistical machine translation with pivot approach was investigated by many researchers. For example Gispert and Mario (2006) used Spanish as a bridge for their Catalan-English translation. They compared two coupling strategies: cascading of two translation systems versus training of system from parallel texts whose target part has been automatically translated from pivot to target. In their work they showed that the phrase translation strategy consistently outperformed the sentence translation strategy in their controlled experiments. Habash and Hu (2009) used English as a pivot language while translating from Arabic to Chinese. Their results showed that pivot strategy outperforms direct translation systems. Babych et al. (2007) used Russian language as a pivot from Ukrainian to English. Their comparison showed that it is possible to achieve better translation quality with pivot approach. Kumar et al. (2007) improved Arabic-English MT by using available parallel data in other languages. Their approach was to combine word alignment systems from multiple bridge languages by multiplying posterior probability matrices. This approach requires parallel data for several languages, like the United Nations or European Parliament corpus. An approach based on phrase table multiplication is discussed in Wu and Wang (2007) .Phrase table is formed for the training process. Scores of the new phrase table are computed by combining corresponding translation probabilities in the source-pivot and pivot-target phrase-tables. They also focused on phrase pivoting. They proposed a framework with two phrase tables: one extracted from a small amount of direct parallel data; and the other extracted from large amounts of indirect data with a third pivoting language. Their results were compared with many different European language as well as Chinese-Japanese translation using English as a pivoting language. Their results show that simple pivoting does not improve over direct MT. Utiyama and Isahara (2007) inspected many phrase pivoting strategies using three European languages (Spanish, French and German). Their results showed that pivoting does not work as well as direct translation. Bertoldi et. al (2008) compare between various approaches of PBSMT models with pivot languages. Their experiments were on Chinese-Spanish translation via disjoint or overlapped English as pivot language. We believe that we are the first to explore the Danish-Arabic language pair directly in MT. We also apply pivoting techniques on none parallel text corpora.

3. System Description

In our work we develop two base lines for each experiment, Danish English and English Arabic. Translation direction is from Danish to Arabic. Moses[1] package is used for training the base lines. The system partition the source sentence into phrases. Each phrase is translated into a target language phrase. We use GIZA++ Och and Ney (2003) for word alignment.

[1]: Moses Package http://www.statmt.org/moses/

We use Pharaoh System suite to build the phrase table and decode (Koehn, 2004). Our language models for both systems were built using the SRILM toolkit Stolcke(2002).We use a maximum phrase length of 6 to account for the increase in length of the segmented Arabic. Our distortion limit set to 6. And finally we use BLEU metric Papineni et al. (2001) to measure performance.

4. Pivot Strategy

We use the phrase-based SMT system described in the previous section to deploy our pivot methods. We inspect two pivot strategies *phrase translation* and *sentence translation*. In both strategies we use English as the pivot language. Danish and Arabic represent source and target languages. In phrase translation strategy we directly construct a Danish-Arabic phrase translation table from a Danish-English and an English-Arabic phrase-table. In sentence translation strategy we first translate a Danish sentence into *n* English sentences and translate these *n* sentences into Arabic separately. We select the highest scoring sentence from the Arabic sentences.

4.1 Sentence Translation Experiment

The sentence translation strategy uses two independently trained SMT systems: a direct Danish-English system and a direct English-Arabic system. We translate every Danish sentence *d* into *n* English sentences *e* {e_1, e_2, ..., e_n} using a Danish-English SMT system. Then we translate each *e* sentence into Arabic sentences *a* {a_1, a_2,..,a_n}. We estimate sentence pair feature according to formula 1 below.

$$\mathrm{S}(\mathrm{s},\mathrm{t}) = \sum_{\mathrm{n}=1}^{8}(\alpha\mathrm{s}_\mathrm{n}\ \beta\mathrm{s}_\mathrm{n}\ + \alpha\mathrm{t}_\mathrm{n}\ \beta\mathrm{t}_\mathrm{n}) \quad ..\ 1$$

$\alpha s_n\ \beta s_n$, $\alpha t_n\ \beta t_n$ is the feature functions for the source and target (s, t) sentences respectively. Feature functions represents: a trigram language model probability of the target language, two phrase translation probabilities (both directions), two lexical translation probabilities (both directions), a word penalty, a phrase penalty, and a linear reordering penalty. Further details on these feature functions is found in (Koehn, 2004; Koehn et al., 2005). We choose to limit the number of the translation for any Danish sentence to English into three due to performance issues. We pass the translation with maximum feature score as input to the English-Arabic system.

4.2 Phrase Translation Experiment

In the phrase translation strategy we need to construct a phrase table to train the phrase-based SMT system. We need a Danish-English phrase table and an English-Arabic phrase-table. From these tables, we construct a Danish-Arabic phrase table. We use a matching algorithm that identifies parallel sentences pairs among the tables. This process is explained in Munteanu and Marcu (2005). We identify candidate sentence pairs using a word-overlap filter tool[1]. Finally we use a classifier to decide if the sentences in each pair are a good translation for each other and update our Danish-Arabic phrase table with the selected pair.

4.3 Data

Data collection was a great challenge for this experiment. Our data resources are from two groups; Arabic-English and English-Danish. Table 1 shows a brief description of our data resources. English-Arabic corpora domain intercrosses with the English-Danish corpora domain to some reasonable degree.

Name	**Direction**	**Domain**	**Size (words)**
Acquis	Danish-English	Legal issues / News	7.0 M
UN multilingual corpus	Arabic-English	Legal issues / News	3.2 M
Meedan	Arabic-English	News	0.5 M
LDC2004T17	Arabic-English	News	0.5 M

Table 1: Corpus resources

	Sample	**Lines**	**Words**
Training	Small	30 K	1 M
	Medium	70 K	2 M
	Large	100 K	3 M
Test	Test (Parallel)	1 K	19 K

Table2: Training and testing data sizes

For the Arabic English we selected three major resources, the United Nations (UN) multilingual corpus[2] which is available at the UN web site.

[1] JTextPro http://sourceforge.net/projects/jtextpro/

[2] UN Corpus http://www.uncorpora.org/

It enjoys a good quality of translation and it contains about 3.2 M lines of data and about 7 M words. The second resource was Meedan[1] corpus, which is a newly developed Arabic English corpus mainly compiled from the internet and news agencies, it contains more than 0.5 M Arabic words. The third resource was provided by LDC [2] (catalog no. LDC2004T17), it contains more than 0.5 M words, it also cover news domain. For the English Danish category we selected the Acquis[3] Corpus, it contains more than 8 K documents and more than 7 M words. Acquis contain many legal documents that cover many domains. English Arabic resources were extracted and aligned using Okapi[4] translation memory editor. With the Acquis corpus we used the available tools that are available at the Acquis website for extracting and aligning Danish English text. All data were tokenized and lowercased separately. In order to inspect the size factor on our SMT system data were compiled into three sets: Large, Medium and Small. Table 2 illustrates the training data size for each set. For testing data we collected a parallel Arabic-English-Danish text from the UN Climate Change conference 2009 which was held in Copenhagen[5]. We extracted 1 K sentences for each language. Table 2 illustrates the training data size for each experiment. The English Arabic corpora domain intercrosses with the English Danish corpora domain to some reasonable degree. We are aware that there might be some bias among data resources coverage, but due to data availability our corpora can still serve our experiments objectives. Given the expense involved in creating direct Arabic-Danish parallel text and given the large amounts of Arabic-English and English-Danish data, we think our approach in collecting data for our experiment is still valid and interesting.

5. Results and Evaluation

We measure our system performance using BLEU scores Papineni et al. (2002). We compare our system performance with Google Translate web service. Comparison with Google provides us with a general performance indicator for our system. Table 3 presents our direct translation system results for DA-EN and EN-AR baselines. As expected BLEU scores will increase when we increase the training data size. We use the same testing data described in section 4.3 with Google Translate; results are described in Table 4. Google outperforms our direct system results especially for the EN-AR direct translation

Training Data Size	**DA-EN**	**EN-AR**
Small	20.3	25.1
Medium	21.4	26.3
Large	23.1	27.1

Table3: BLEU Scores for Direct Sentence Based SMT systems.

Our direct system for DA-EN system BLEU score was 23 which is (64%) of Google system BLEU scores while for the EN-AR system BLEU score was 27.1 which is (40%) of Google system BLEU scores.

	DA-EN	**EN-AR**	**DA-AR**	**DA-EN-AR**
Test Sample	36.0	67.0	30.0	30.0

Table 4 describe the BLEU scores for Google translate web service on our test sample

In Table 5 we present the results of the sentence pivoting system and the phrase pivoting system. Sentence based strategy outperform Phrase based strategy. For the large size training data set the system achieved a score of 19.1 for the sentence based system compared with 12.9 to the phrased based strategy .This results differs from previous similar studies like Utiyama and Isahara (2007) and Habash and Hu (2009) where pivot strategy outperform sentence strategy. Pivot system was not better because of the quality and quantity of the DA-EN-AR phrase table entries which was received from the matching algorithm. Pivot system is dependent on the matching algorithm and enhancing it will enhance system performance. Google DA-EN and DA-EN-AR results were the same. This is a good indicator that Google uses pivot approach between languages with limited resources like the case of Arabic and Danish. Figure 1 represents a sample of our best performing system results, compared with Google translate web service. The sample shows both original text and its translation, and our system translation results for the same text.

[1] Meedan http://github.com/anastaw/Meedan-Memory
[2] LDC http://www.ldc.upenn.edu/
[3] Acquis http://langtech.jrc.it/JRC-Acquis.html
[4] Okapi http://okapi.sourceforge.net/
[5] Cop15 http://en.cop15.dk

Size	Sentence Based Pivot Strategy (Da- En- Ar)	Phrase Based Pivot Strategy (Da- En - Ar)
Small	15.0	11.4
Medium	16.9	12.3
Large	19.1	12.9

Table5: BLEU Scores for Phrase based and Sentence Based SMT systems.

6. Conclusion and Future work

Developing a SMT system between two language pairs that don't share many linguistic resources Like Danish and Arabic language pairs is a quite challenging task. We presented a comparison between two common pivot strategies; phrase translation and sentence translation. Our initial results show that sentence pivot strategy outperforms phrase strategy especially when common parallel corpora are not available. We compared our system results with Google translate web service to estimate relative progress and results were promising. In the future we plan to enhance our pivoting techniques. Phrase pivot strategy is still a promising technique we need to utilize with our baseline. Phrase Pivot strategy performs better when more parallel data resources are available, so we plan to collect more parallel training data for our baseline. We also plan to apply state of the art alignment technique and to use word reorder tools on our system training data. This will enhance our SMT system learning process. We also plan to train our SMT system to fit domain specific areas like weather, or climate domains. We target high quality pivot techniques that will help us outperform available commercial tools like Google Translate especially for domain specific SMT areas

Reference	**DA**	Jeg tror, at en af de store mangler ved Kyoto var, at den officielle delegation kom tilbage med en aftale, som de vidste aldrig ville blive vedtaget i senatet.
	EN	I think that a major shortcoming of Kyoto was that the official delegation came back with a treaty they knew was never going to make it through the Senate
	AR	وأعتقد أن أحد أوجه القصور الرئيسية في كيوتو هو أن الوفد الرسمي عاد مع معاهدة كانوا على علم أنها لن تمر خلال مجلس الشيوخ
System		وأعتقد أن أحد مشاكل الرئيسيةفي Kyoto كان الوفد الرسمي جاء يعود مع أنهم اعلم كان لن يتم اعتماده
Google		اعتقد ان احد العيوب الرئيسية في كيوتو هو أن الوفد الرسمي عاد الى اتفاق مع أنهم يعرفون لن يتم اعتماده في مجلس الشيوخ
Reference	**DA**	Men selv om udledningen af drivhusgasser forventes at falde på grund af faldende aktivitet i industrien, tror de Boer ikke, det vil mindske presset på landene om at handle og underskrive en ny aftale.
	EN	But even though greenhouse gas emissions are expected to slow down as a result of shrinking industrial activities ,de Boer does not believe it will lessen the pressure on countries to act and sign a new treaty.
	AR	و على الرغم من الانبعاثات الغازية لبيت الدفيئة من المتوقع أن تنخفض نتيجة لانخفاض الأنشطة الصناعية ، دي بوير لا يعتقد أن ذلك سوف يقلل من الضغط على الدول للعمل والتوقيع على معاهدة جديدة
System		حتى على الرغم من انبعاثات غازات الحرارة من المتوقع تنخفض على أاس النشاط التنازلي,وستحدد الضغوط على البلاد لعمل على الاتفاقية جديدة .

Figure 1: Selected samples of system translation result

References

A. de Gispert and J. B. Mario, “Catalan-english statistical machine translation without parallel corpus: bridging through spanish,” in Proc. of 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 2006.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In ICSLP.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from underresourced languages: comparing direct transfer against pivot translation. In Proceedings of MT Summit XI, Copenhagen, Denmark.

Callison-Burch et al. (2006) Chris Callison-Burch, Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. InIWSLT.

Callison-Burch et al (2006)Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In Proceedings of HLT-NAACL’06. New York, NY, USA

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51.

Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. In Proceedings of Coling ACL’06. Sydney, Australia

H. Wu and H. Wang, “Pivot language approach for phrase-based statistical machine translation,” in Proc.of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic,2007, pp. 856–863.

Ibrahim Badr, Rabih Zbib, and James Glass 2008. Syntactic Phrase Reordering for English-to-Arabic Statistical Machine Translation. In Proc. of ACL/HLT.

Jakob Elming,2008, Syntactic Reordering Integrated with Phrase-based SMT , ACL proceedings.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In ACL.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In Proceedings of NAACL-HLT’07, Rochester, NY, USA

Munteanu and Marcu (2005)Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics,31(4):477–504.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, Roldano Cattoni, 2008 ,Phrase-Based Statistical Machine Translation with Pivot Languages, Proceedings of IWLST , USA.

Nizar Habash and Jun Hu.2009. Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 173–181.

Philipp Koehn, Alexandra Birch and Ralf Steinberger: 462 Machine Translation Systems for Europe, in Proceedings of the 12th MT Summit, (Ottawa, Canada, 26-30 August, 2009), p. 65-72.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. Computational Linguistics,29(3):349–380.

Shankar Kumar, Franz Och, and Wolfgang Macherey.2007. Improving word alignment with bridge languages. In Proceedings of EMNLPCoNLL’ 07, Prague, Czech Republic.

Yaser Al-Onaizan and Kishore Papineni. 2006 Distortion models for statistical machine translation.In Proceedings of Coling-ACL’06. Sydney, Australia.